A Project Report On

Analysis of Global Average Temperature Using ARIMA and LSTM Models

Submitted in partial fulfillment of the requirement for the 4^{th} semester

Master of Science
in
STATISTICS

DEPARTMENT OF STATISTICS,
SCHOOL OF MATHEMATICAL SCIENCES,
KAVAYITRI BAHINABAI CHAUDHARI NORTH MAHARASHTRA UNIVERSITY JALGAON - 425001

Submitted By

Mr. Marathe Harshal Balkrushna (Seat no.- 358783)

Ms. Patil Komal Sitaram (Seat no.- 358788)

Mr. Patil Ratnesh Dnyaneshwar (Seat no.- 358792)

Under the guidance of

Prof. R. L. Shinde
Senior Professor, Department of Statistics

(2024–2025)

CERTIFICATE

This is to certify that **Mr. Harshal Balkrushna Marathe, Ms.Komal Sitaram Patil and Mr. Ratnesh Dnyaneshwar Patil** students of M.Sc. (Statistics), at Department of Statistics ,Kavayitri Bahinabai Chaudhari North Maharashtra University, Jalgaon have successfully completed their project work entitled **“Analysis of Global Average Temperature Using ARIMA and LSTM Models”** as a part of M.Sc. (Statistics) program under my guidance and supervision during the academic year 2024-2025(Sem-IV).

Prof. R. L. Shinde

Acknowledgement

We are pleased to have successfully completed the project **“Analysis of Global Average Temperature Using ARIMA and LSTM Models”**. We thoroughly enjoyed working on this project and gained valuable knowledge throughout the process.

We would like to express our heartfelt gratitude to **Prof. K. K. Kamalja**, Head of the Department of Statistics, for granting us permission to use institutional facilities.

We are deeply thankful to our respected guide, **Prof. R. L. Shinde**, Professor, Dept. of Statistics, for his continuous guidance, encouragement, and insightful suggestions throughout the project.

We also extend our sincere thanks to **Asst. Prof. M. C. Patil** and **Asst. Prof. R. D. Koshti** for their help and support.

We appreciate all the faculty members and support staff of the Department of Statistics, School of Mathematical Sciences. Without their support, this work would not have been possible.

Finally, we thank our classmates and families for their constant support and motivation during this academic journey.

Harshal Marathe

Komal Patil

Ratnesh Patil

Abstract

This study analyzes global temperature anomalies over a period of 140 years (1880–2020) using a combination of statistical methods and machine learning techniques. The primary aim is to understand the long-term trends, growth rates, and forecasting of global temperature anomalies. Statistical analysis, including the calculation of monthly, yearly, and multi-year averages, as well as significance tests (t-tests and ANOVA), reveals a clear upward trend in temperature anomalies, supporting the hypothesis of global warming. Distribution fitting indicates a skewed distribution of temperature anomalies, reinforcing the trend toward higher positive anomalies.

Time series analysis using ARIMA models (1,1,4) captures the underlying trends and seasonal variations in the data, with the ARIMA model showing good forecasting performance based on accuracy metrics such as MAE, RMSE, and AIC. Additionally, Long Short-Term Memory (LSTM) models, a form of deep learning, are employed to capture non-linear relationships and improve forecasting accuracy. The LSTM model significantly outperforms the ARIMA model, demonstrating its ability to capture the complex dependencies and long-term forecasting capabilities of global temperature anomalies.

This study highlights the potential of machine learning models, particularly LSTM, in improving the accuracy of long-term climate predictions. The research also suggests that combining traditional statistical models with advanced machine learning techniques could provide more robust forecasts and valuable insights into future climate change scenarios. The findings underscore the urgency of addressing global warming, emphasizing the accelerating rate of temperature rise in recent decades.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background

Climate change has been a pressing global concern, with rising temperatures, extreme weather events, and environmental shifts signaling a long-term warming trend. Recent reports indicate that 2024 was the first calendar year to breach the 1.5°C threshold above pre-industrial levels. The implications of this milestone extend to climate policies, mitigation strategies, and predictive modeling efforts. This section contextualizes our study by referring to recent developments and their scientific relevance.

1.1.1 Global Average Temperature Rise: 2024 Breach of the 1.5°C Threshold

According to the Copernicus Climate Change Service (ECMWF) and the World Meteorological Organization (WMO), 2024 became the first year to exceed an annual average temperature anomaly of 1.5°C above pre-industrial levels (1850-1900 average). While this breach does not indicate the complete failure of the Paris Agreement targets—since the agreement refers to long-term trends over decades—it highlights the accelerating pace of global warming The Indian Express (2025a). Several factors contributed to the record-breaking temperatures of 2023 and 2024. The presence of an El Niño event amplified global temperatures, but additional influences, such as decreased sulfur dioxide emissions from shipping and potential impacts from the solar cycle, may have also played a role. The increasing rate of warming, currently over 0.2°C per decade, suggests that the 1.5°C target could be exceeded permanently within the 2030s.

1.1.2 Melting Glaciers and Rising Sea Levels

A study published in Nature on February 19, 2025, reported that glacier melt has led to nearly 2 cm of sea level rise since the beginning of the 21st century. While this figure may seem small, every centimetre of sea level rise puts an additional two million people at risk of annual flooding. The combination of glacier mass loss and thermal expansion of seawater is accelerating sea level rise, with rates now reaching 0.42 cm per year The Indian Express (2025b), more than doubling since the early 1990s. Sea level rise is not uniform across the globe. Mumbai, for example, has experienced a 4.44 cm increase between 1987 and 2021.

Other coastal cities, such as Visakhapatnam and Kochi, are also facing rising threats. As a result, coastal flooding, ecosystem disruptions, and infrastructure challenges are becoming more frequent, with significant socio-economic consequences.

1.1.3 Decline in Global Sea Ice Cover

In February 2025, the combined Arctic and Antarctic Sea ice extent dropped to a record low of 15.76 million sq km, breaking the previous low recorded in early 2023. While Arctic Sea ice has been shrinking for decades, the Antarctic has recently shown significant losses as well, reversing a long-standing trend of stability The Indian Express (2025c). The loss of sea ice has multiple implications:

1. Reduced ice cover leads to higher oceanic heat absorption, further accelerating global warming.
2. Disruptions in ocean circulation patterns due to freshwater influx from melting ice can impact global weather systems.
3. Thinner and more fragmented ice makes Arctic and Antarctic regions more susceptible to storm-induced breakups.

These findings align with the ongoing temperature rise trends and provide additional motivation for studying global temperature anomalies using statistical and machine learning models. The insights from this study could aid in understanding historical trends, improving predictions, and informing climate policy.

Summary of Key Events and Sources

- January 13, 2025: Reports confirm 2024 as the first full year to breach the 1.5°C threshold.
- February 24, 2025: Nature study highlights the significant impact of glacier melt on global sea level rise.
- February 20, 2025: NSIDC and BBC analysis report record-low sea ice coverage in both polar regions.

These recent developments reinforce the urgency of climate change research, making our study on global temperature anomalies both timely and essential.

1.2 Climate Change and the Importance of Temperature Anomaly Studies

Climate change has emerged as one of the most pressing global challenges of the 21st century. Rising global temperatures, increasing frequency of extreme weather events, and shifts in climate patterns have profound implications for ecosystems, economies, and human societies. Understanding the extent and causes of climate change is essential for developing mitigation and adaptation strategies. One of the key indicators of climate change is the variation in surface temperatures over time. However, absolute temperature measurements vary widely due to geographical and seasonal differences. To overcome this challenge, temperature anomalies are used instead of raw temperature values. A temperature anomaly represents the deviation of a measured temperature from a long-term average reference period. This standardized approach allows for meaningful comparisons across different regions and time periods.

1.3 Motivation

Recent climate reports have highlighted unprecedented environmental changes, reinforcing the urgency of climate change research. The confirmation that 2024 was the first full year to exceed the 1.5°C warming threshold (January 13, 2025) signals a critical moment in global climate history. Additionally, new research published in Nature (February 24, 2025) emphasizes the accelerating contribution of glacier melt to sea level rise, while reports from NSIDC and BBC (February 20, 2025) confirm record-low sea ice coverage at both poles. These alarming developments underscore the necessity of understanding long-term temperature trends and their broader impacts on Earth's climate system. Motivated by these findings, our study aims to provide a comprehensive Global Warming: A Comprehensive Study of Temperature Anomalies Using Statistical and Machine Learning Model anomalies using the Berkeley Earth dataset. By employing advanced statistical techniques and machine learning models, we seek to refine our understanding of historical temperature variations, detect key climate shifts, and assess the reliability of different temperature datasets. Through this research, we hope to contribute valuable insights into the mechanisms driving global warming and support informed decision-making for climate action. The following sections outline the data sources, methodology, statistical analysis, and key findings derived from this study.

1.4 Sustainable Development Goal (SDG) 13 :Climate Action and Global Average Temperature Trends

This research aligns with the United Nations Sustainable Development Goal (SDG) 13, which emphasizes urgent action to combat climate change and its impacts. Specifically, our study contributes to the following SDG 13 targets:

1. Strengthening resilience and adaptive capacity to climate-related hazards by identifying trends in extreme temperature events and their effects on ecosystems and human societies.

2. Integrating climate change measures into national policies and planning by providing robust statistical evidence on global temperature anomalies that can inform mitigation and adaptation strategies.

3. Enhancing education, awareness, and institutional capacity on climate change through the dissemination of findings and development of predictive models for temperature anomalies.

4. Supporting climate action by contributing to the scientific understanding of temperature trends, which can help drive policy and funding decisions toward climate mitigation efforts.

5. Promoting capacity-building in least developed countries by providing accessible data analysis methodologies and encouraging informed decision-making strategies.

1.5 Research Questions

This study seeks to answer the following key research questions:

- What are the long-term trends in global average temperature across different time periods?
- How do global average temperature vary seasonally and Periodically ?
- What statistical techniques best capture historical temperature trends and predict future global average temperature?
- How do machine learning models (e.g., LSTMs) compare to traditional statistical models (e.g., ARIMA) in forecasting global average temperature ?
- What role do extreme temperature events play in the observed global average temperature , and how can they be modeled?

1.6 Objective

The primary aim of this study is to analyze global average temperature using statistical and machine learning models. The specific objectives include:

1. **Study Temperature Anomalies Over Time** – Examine temperature changes over different time periods, such as yearly, every 10 years, and long-term trends.
2. **Identify Patterns and Trends** – Find patterns, trends, and seasonal changes in global temperatures to understand how climate is changing.
3. **Analyze Temperature Distributions** – Study how temperature anomalies are spread out and change over time.
4. **Calculate Growth Rates** – Measure how much global temperatures have increased or decreased over different time periods.
5. **Find Trends, Seasonality, and Residuals** – Separate the data into trends (long-term changes), seasonality (repeating patterns), and random variations.
6. **Apply Time Series Models** – Use statistical models like ARIMA, exponential smoothing, and trend analysis to study past temperature changes.
7. **Evaluate Machine Learning Models** – Test machine learning methods, such as LSTM, to see how well they can predict temperature changes.
8. **Compare Statistical and Machine Learning Models** – Compare traditional statistical methods with machine learning approaches to find the best model for temperature forecasting.
9. **Develop Predictive Models** – Create models that can predict future temperature changes in both the short term and the long term.
10. **Study Climate Feedback Effects** – Investigate factors that speed up or slow down global warming.

11. **Use Findings for Climate Policies** – Use the study's results to help policymakers make informed decisions about climate change and adaptation strategies.

By addressing these objectives, this study seeks to advance the understanding of global temperature anomalies and contribute to the broader discourse on climate change mitigation and adaptation.

1.7 Scope of the Project

This study utilizes the Berkeley Earth daily TAVG full dataset, which provides global land-surface temperature anomalies. The dataset covers the period 1880–2020, with anomalies reported relative to the 1951–1980 average temperature baseline. The key characteristics of the dataset include:

- Temporal Coverage: 1880–2020
- Daily observations : 52078
- Measurement Unit: Temperature anomalies in Celsius
- Reference Baseline: Mean temperature from January 1951 – December 1980 (8.59°C ± 0.05°C)

This research aims to analyze long-term temperature trends, examine seasonal variations, and evaluate prediction accuracy using statistical and machine learning techniques. The study employs time series models (ARIMA, LSTMs, etc.) to understand temperature fluctuations and forecast future anomalies.

Limitations
While the dataset is valuable for understanding global temperature trends, a few important limitations should be noted:

- **No Country-Level Data:**
  - The dataset shows global patterns but does not include country-specific temperature values in a downloadable format.
- **Missing or Less Reliable Data in Early Years:**
  - Some old records (especially before 1900) may be less accurate due to fewer measurements and inconsistent methods.
- **Reference Period May Affect Results:**
  - The anomalies are based on the 1951–1980 average. This may not fully reflect more recent climate changes.

Despite these limitations, the dataset remains a strong foundation for studying global warming and building predictive climate models.

1.8 Literature Review

Understanding changes in global temperature is important for studying climate change and its effects on our planet. Many researchers have studied historical temperature data to find patterns and causes of global warming. This section reviews important studies on global temperature trends, methods used for analysis, and key findings from climate research.

Studies on Global Temperature Trends

Global Temperature Evolution 1979–2010

The study *Global Temperature Evolution 1979–2010* by Foster and Rahmstorf (2011b) presents a comprehensive analysis of global temperature trends using data from five prominent datasets—three surface temperature records (NASA/GISS, NOAA/NCDC, HadCRUT) and two lower-troposphere records (RSS and UAH). This research builds upon the foundational work of prior studies that have identified the patterns and causes of global warming.

Previous investigations, such as Lean and Rind (2008, 2009), have established a framework for understanding the influence of anthropogenic factors, volcanic activity, solar variability, and the El Niño-Southern Oscillation (ENSO) on global temperatures. These studies revealed the dominant role of human-induced factors in driving long-term warming trends while highlighting the significant impact of natural short-term variations on temperature records.

Foster and Rahmstorf contribute to this discourse by focusing on the period since 1979, when satellite data became available, allowing for a linear approximation of warming trends. Their methodology, which utilizes multiple regression analysis to remove the influence of ENSO, volcanic aerosols, and solar irradiance, aligns with earlier approaches but offers increased precision and robustness. By doing so, they effectively isolate the global warming signal, emphasizing its steady progression over the 32-year study period.

The study corroborates findings from Hansen et al. (2010) regarding the consistent warming of both surface and lower-troposphere temperatures. However, it also highlights discrepancies between these datasets, particularly in their response to exogenous factors—a nuance noted in earlier works by Mears and Wentz (2008) and Christy et al. (2000).

While this research reinforces the urgency of addressing anthropogenic climate change, it also underscores the challenges inherent in analyzing temperature records, given the variability induced by natural factors. The study's conclusions are a testament to the advances in climate science, yet they call for continued refinement in data processing methodologies to ensure accuracy in future assessments.

NASA GISS Temperature Study (Hansen et al., 2010)

The study by Hansen et al. (2010) updates the Global Warming: A Comprehensive Study of Temperature Anomalies Using Statistical and Machine Learning Model changes by NASA's Goddard Institute for Space Studies (GISS). It compares different methods of analyzing temperature data and addresses common misunderstandings about global warming.

The researchers used satellite images of night lights to identify weather stations in very dark areas. They adjusted temperature data from cities and surrounding areas to remove errors caused by human activities, confirming that cities have little impact on overall global temperature trends. The study combined ocean temperature records with land-based weather station data and tested different ways of estimating ocean temperatures. It found that global temperature trends are affected by how temperatures in the polar regions are estimated since there are fewer observations in those areas.

The researchers also used 12-month averages to make temperature trends clearer in their graphs. They found that, despite yearly fluctuations, the rate of global warming has not slowed down. The Earth's temperature has been increasing just as fast in the last decade as in the previous two decades. In fact, the highest global temperature recorded in the study period was reached in 2010.

Global Temperature Change

The research letter *Global Temperature Change* by Hansen et al. (2006) provides an analysis of global surface temperature trends and emphasizes the role of human-induced greenhouse gas emissions in recent warming. Building on earlier work by Manabe and Wetherald (1975) and Hansen et al. (1988), this study underscores the correlation between anthropogenic activities and climate change.

The study highlights the rapid rate of warming in recent decades, surpassing natural variability observed in the Holocene epoch. Hansen et al. (2006) identify human activities, such as fossil fuel burning and deforestation, as dominant drivers of this trend. They also note regional temperature variations, including disproportionate warming in the Western Equatorial Pacific, which may influence El Niño events.

By integrating paleoclimate evidence, Hansen et al. argue that current global temperatures are approaching levels unseen for millennia. The authors call for limiting global warming to 1°C above year-2000 levels to avoid dangerous climate impacts, aligning with the recommendations of the Intergovernmental Panel on Climate Change (IPCC).

NOAA Climate Report (Lindsey & Dahlman, 2020)

Lindsey and Dahlman (2020), in their NOAA climate report, analyzed global temperature changes over the past 100 years. They found that Earth's surface temperature has increased significantly, mostly due to human activities like burning fossil fuels. The study showed that the 10 hottest years ever recorded have all happened in the past two decades, proving that global warming is getting worse. The report also highlighted how rising ocean temperatures are changing weather patterns, raising sea levels, and affecting marine life.

Long-Term Temperature Changes (Foster & Rahmstorf, 2011)

Foster and Rahmstorf (2011a) studied how natural events like volcanic eruptions, El Niño, and solar activity affect temperature changes. They removed these short-term effects from the data and found that human activities, especially greenhouse gas emissions, are the main cause of rising global temperatures. Even after removing natural factors, the overall warming trend remained strong.

Methods Used to Study Global Temperature

Traditional Statistical Methods

Many researchers use statistical models to analyze temperature trends. Methods like ARIMA (AutoRegressive Integrated Moving Average) and exponential smoothing help in predicting future temperatures. Hansen et al. (2010) used a method called the 12-month running mean to smooth out short-term changes and focus on long-term temperature trends.

Machine Learning Models

Recently, scientists have started using machine learning methods like Long Short-Term Memory (LSTM) networks to predict temperature changes. These models can detect complex patterns in climate data and make more accurate forecasts. Researchers compare machine learning models with traditional statistical models to find out which one works best for temperature predictions.

Effects of Climate Change on Temperature

- **Impact of Greenhouse Gases**
  Many studies have shown that rising carbon dioxide (CO_2) levels are directly linked to increasing global temperatures. The NOAA report explained that human activities, such as burning coal, oil, and gas, have increased the amount of CO_2 in the atmosphere, trapping more heat and causing long-term warming.

- **Temperature Changes in the Polar Regions and Oceans**
  Hansen et al. (2010) and Lindsey and Dahlman (2020) pointed out that temperature changes in the Arctic and Antarctic play a major role in global climate patterns. These regions are warming faster than the rest of the world, causing ice to melt, sea levels to rise, and weather patterns to shift. In Research of Rignot et al. (2019) highlights the phenomenon of "polar amplification," where the Arctic and Antarctic regions are warming at rates significantly higher than the global average. This effect is driven by feedback mechanisms, such as the albedo effect, where melting ice reduces the Earth's reflectivity, causing further warming. Hansen et al. (2006) emphasize that the Arctic is warming twice as fast as the rest of the planet, resulting in shrinking sea ice, thawing permafrost, and habitat loss for Arctic species.

  In the Antarctic, the impacts are regionally varied. The Antarctic Peninsula is warming rapidly, contributing to ice shelf disintegration and glacier retreat, while East Antarctica shows more resilience. Rignot et al. (2019) highlight the vulnerability of West Antarctica's ice sheets, which are contributing significantly to global sea level rise due to increased ice flow into the ocean.

  The surrounding oceans are also undergoing rapid changes. Paleoclimate studies suggest that current temperatures in the polar regions are approaching levels not seen for millennia. The combination of modern observations and historical data underscores the urgency of reducing greenhouse gas emissions to prevent further destabilization of these vital regions.

Chapter 2

Data and Methods

2.1 Introduction

This chapter outlines the data sources, preprocessing steps, and modeling techniques employed in the analysis. It describes the characteristics of the dataset, justifies the preprocessing techniques applied, and details the statistical and machine learning methods used for modeling and evaluation.

2.2 Data Description

This study utilizes temperature anomalies instead of raw temperature data to address several analytical challenges:

- Geographical Variations: Different regions experience vastly different climate conditions (e.g., tropical vs. polar climates), making absolute temperature comparisons misleading.
- Seasonal Variations: Temperature fluctuates throughout the year, and anomalies help in removing these variations to allow year-to-year comparisons.
- Uneven Distribution of Weather Stations: Some areas have dense temperature monitoring networks, while others (e.g., oceans, polar regions) have sparse or incomplete data.

By using anomalies, we eliminate these biases and focus on temperature deviations relative to a common historical baseline. The anomalies in this study are computed relative to the January 1951 – December 1980 mean temperature, allowing for consistent comparisons across different time periods. For example, if the average January temperature during the reference period was 10°C, and a given January in 2000 recorded 12°C, the anomaly would be +2°C. This standardized approach facilitates the analysis of global climate trends over time.

2.2.1 Source of Data

This study utilizes **secondary data** obtained from the **Berkeley Earth Global Temperature Dataset**, which provides a comprehensive record of global land–ocean temperature anomalies. This dataset has been constructed by combining the Berkeley Earth monthly land temperature fields with a spatially kriged version of the HadSST3 sea surface temperature dataset. The resulting product spans the period from 1850 to the present and offers extensive global coverage, with approximately:

- 57% of Earth's surface covered in 1850
- 75% coverage by 1880
- 95% coverage by 1960
- 99.9% coverage by 2015

Temperature data is provided in a regular $1^\circ \times 1^\circ$ latitude–longitude grid, with monthly averages for each grid cell when available. The dataset includes two versions that differ in how regions with sea ice are treated: either as air temperature over sea ice (preferred for most applications) or as sea surface temperature beneath sea ice. This distinction significantly impacts temperature anomaly trends due to rapid warming in the Arctic. Incorporating Arctic air temperature changes suggests an additional global-average warming of approximately 0.1°C since the 19th century when compared with datasets that neglect this adjustment.

The Berkeley Earth dataset is comparable to other prominent temperature series such as:

- Hadley's HadCRUT4
- NASA's GISTEMP
- NOAA's GlobalTemp
- Cowtan and Way

This dataset offers a spatially complete and homogeneous temperature field, making it highly suitable for long-term climate change analysis.

2.2.2 Berkeley Earth Dataset Features

The Berkeley Earth dataset is renowned for its comprehensive coverage and methodological robustness. According to Rohde and Hausfather (2020a), the Berkeley Earth dataset provides a spatially complete and homogeneous global temperature record. Developed through peer-reviewed processes, it incorporates a larger volume of temperature observations than many other datasets, ensuring improved spatial representation and accuracy.

- **Global Monthly Averages (1850 – present)**: Combines land and ocean records for long-term climate trend analysis.
- **Global Land-Only Data (1750 – present)**: Tracks temperature anomalies across land regions.
- **Global Daily Land Temperature (1880 – present)** [Experimental]: Provides daily variations for high-resolution analysis.
- **Gridded Temperature Data**: Available in equal-area and lat–long formats, facilitating regional studies.

A unique aspect of the Berkeley Earth methodology is the use of machine learning techniques to enhance spatial resolution and interpolate missing data in regions with sparse measurements. The dataset is particularly useful for climate and environmental research due to the following advantages:

- Extensive time span covering over 140 years
- Advanced bias correction techniques
- Seamless integration of land and ocean temperature records

The dataset is regularly updated and available at the Berkeley Earth website: `http://berkeleyearth.org/data/`. A version of the dataset discussed in this project has been archived and can be accessed via Zenodo: `https://doi.org/10.5281/zenodo.3634713` Rohde and Hausfather (2020b).

2.2.3 Sample Data Snapshot

Table 2.1: **Sample snapshot from the global average temperature dataset (1880–2020)**

Year	Month	Day	Day of Year	Anomaly (°C)	Temperature (°C)
1880	1	1	1	-0.692	7.898
1880	1	2	2	-0.592	7.998
1880	1	3	3	-0.673	7.917
1880	1	4	4	-0.615	7.975
1880	1	5	5	-0.681	7.909
1880	1	6	6	-0.743	7.847
1880	1	7	7	-0.646	7.944
⋮	⋮	⋮	⋮	⋮	⋮
2020	12	25	359	0.644	9.234
2020	12	26	360	0.462	9.052
2020	12	27	361	0.585	9.175
2020	12	28	362	0.796	9.386
2020	12	29	363	0.840	9.430
2020	12	30	364	0.815	9.405
2020	12	31	365	0.705	9.295

2.2.4 Data Attributes

The dataset consists of the following attributes:

- **Year** – The calendar year (e.g., 1880, 1881, ...).
- **Month** – Month of the year (1 to 12).
- **Day** – Day of the month (1 to 31).
- **Day of Year** – Day number within the year (1 to 365).
- **Anomaly** – Temperature anomaly relative to a baseline period.
- **Temperature** – Actual recorded global temperature.

2.2.5 Temperature Anomalies and Reference Period

What are Temperature Anomalies?

A **temperature anomaly** is the difference between an observed temperature and a reference or baseline temperature. Instead of using absolute temperatures, anomalies provide a clearer picture of long-term temperature trends by removing local variability.

Mathematically, a temperature anomaly is calculated as:

$$\text{Anomaly} = \text{Observed Temperature} - \text{Reference Temperature} \tag{2.1}$$

where the **Reference Temperature** is the average temperature over a fixed period.

Why Use Anomalies Instead of Absolute Temperatures?

Temperature anomalies are preferred over absolute temperatures because:

- They allow for better **comparisons across different regions**, as absolute temperatures vary widely by location.
- They help **highlight long-term climate trends** more effectively.
- They reduce **bias due to changes in measurement stations** over time.

Reference Period (1951-1980)

The reference period used in this study is **1951-1980**, which is commonly adopted by organizations such as NASA, NOAA, and Berkeley Earth. This period is selected because:

- It represents a stable climate period before the rapid warming seen in recent decades.
- It has **good global data coverage**, making it a reliable baseline.
- It allows for consistent **comparisons with modern temperature data**.

By comparing modern temperature anomalies to this reference period, we can analyze how global temperatures have changed over time.

2.3 Data Preprocessing

Missing Data Handling

Missing values were identified using descriptive statistics and visual inspection. Forward fill and linear interpolation methods were used for imputation.

Outlier Detection

Outliers were detected using IQR and visualized through boxplots. Extreme outliers were treated using winsorization and contextual domain-based correction.

Data Smoothing

To reduce noise and highlight trends, a 7-day moving average smoothing technique was applied to the temperature series.

Data Transformation

All temperature and anomaly values were standardized using z-score normalization. Stationarity was enforced through first-order differencing for ARIMA modeling.

2.4 Methodology

2.4.1 Stochastic Analysis of Daily Global Average Temperature Trends

Markov chains provide a fundamental stochastic framework for modeling systems that undergo transitions from one state to another over time. A Markov process is a sequence of random variables where the probability of moving to the next state depends solely on the current state, not on the sequence of states that preceded it. This memoryless property is known as the Markov property.

In the context of climate analysis, Markov chains can be useful in studying directional trends in temperature. Here, we define a two-state process X_n derived from daily global average temperatures:

$$X_n = \begin{cases} 1 & \text{if } T_n \geq T_{n-1} \quad \text{(temperature increase or no change)} \\ -1 & \text{if } T_n < T_{n-1} \quad \text{(temperature decrease)} \end{cases}$$

This binary sequence represents the day-to-day directional trend in temperature and forms the basis of a discrete-time, two-state Markov chain. To verify if the sequence satisfies the Markov property, we examine the transition probabilities:

$$P(1 \to 1), \quad P(1 \to -1), \quad P(-1 \to 1), \quad P(-1 \to -1)$$

These form the Transition Probability Matrix (TPM):

$$P = \begin{bmatrix} P(1 \to 1) & P(1 \to -1) \\ P(-1 \to 1) & P(-1 \to -1) \end{bmatrix}$$

To statistically test the Markov property, we compare observed transition frequencies with expected frequencies under the assumption of independence. A chi-square test can be used to assess whether transitions are dependent on the previous state, thereby supporting the Markov assumption.

Analyzing temperature trends through this lens provides insights into persistence or switching behavior in climate data over time, which can be crucial for understanding long-term global warming patterns.

2.4.2 Trend Analysis in Time Series

Trend analysis involves examining patterns or tendencies in data over a period of time to identify consistent increases, decreases, or stability.

Definition of Trend

A **trend** represents the long-term progression of a time series, typically ignoring short-term fluctuations and seasonal variations.

Importance

- Detects long-term patterns in variables like temperature, stock prices, or sales.
- Assists in forecasting and decision-making.
- Critical in understanding climate change over time.

Common Methods of Trend Analysis

1. **Moving Average:**

$$MA_t = \frac{1}{k}\sum_{i=0}^{k-1} X_{t-i}$$

2. **Linear Regression:**

$$Y_t = \beta_0 + \beta_1 t + \epsilon_t$$

Where β_1 indicates the presence and direction of a trend.

3. **Differencing:** Removes trend to achieve stationarity:

$$Y_t' = Y_t - Y_{t-1}$$

4. **Seasonal Decomposition:** Separates the time series into:

$$Y_t = T_t + S_t + R_t$$

where T_t is the trend, S_t is the seasonality, and R_t is the residual.

2.4.3 Growth Rate Calculation

Growth rate measures how a quantity changes over time and is commonly expressed as a percentage.

Simple Growth Rate

The basic formula to compute the percentage growth rate between two values is:

$$\text{Growth Rate} = \left(\frac{\text{Final Value} - \text{Initial Value}}{\text{Initial Value}}\right) \times 100$$

N yearly Annual Growth Rate

Used to calculate the average rate of change per year.

$$\text{Growth Rate} = \left(\frac{\text{Final Value} - \text{Initial Value}}{\text{Initial Value}}\right)^{1/n} \times 100$$

2.4.4 ARIMA Model Explanation

ARIMA stands for **AutoRegressive Integrated Moving Average**, a popular model used in time series forecasting.

ARIMA(p, d, q)

An ARIMA model is defined by three parameters:

- **p** - the number of autoregressive (AR) terms
- **d** - the number of nonseasonal differences needed for stationarity
- **q** - the number of lagged forecast errors in the prediction equation (MA terms)

Model Components

- **AutoRegressive (AR)**: The model uses the dependency between an observation and a number of lagged observations.
- **Integrated (I)**: Differencing of raw observations is used to make the time series stationary.
- **Moving Average (MA)**: The model uses the dependency between an observation and a residual error from a moving average model applied to lagged observations.

Mathematical Representation

For ARIMA(1,1,1), the model can be written as:

$$y_t = \mu + \phi_1 y_{t-1} + \theta_1 \varepsilon_{t-1} + \varepsilon_t$$

Where:

- y_t: observed value at time t
- μ: constant mean
- ϕ_1: autoregressive coefficient
- θ_1: moving average coefficient
- ε_t: error term (white noise)

Assumptions

- The series should be stationary (or made stationary using differencing).
- Residuals should be white noise.

2.4.5 Hybrid Model: Polynomial Regression + Residual ARIMA

To capture both deterministic trends and stochastic patterns in the temperature anomaly time series, we employed a **hybrid modeling approach** that combines **Polynomial Regression** for trend estimation and **ARIMA modeling** for residual analysis and correction. This method leverages the strengths of both models—regression for capturing nonlinear trends and ARIMA for modeling temporal dependencies in the residuals.

1. Polynomial Regression (Degree 3)

Polynomial regression is an extension of linear regression in which the relationship between the independent variable t (typically time) and the dependent variable Y_t (e.g., temperature) is modeled as an n^{th}-degree polynomial. For a degree-3 polynomial regression model, the equation is:

$$Y_t = \beta_0 + \beta_1 t + \beta_2 t^2 + \beta_3 t^3 + \epsilon_t$$

where:

- $\beta_0, \beta_1, \beta_2, \beta_3$ are the regression coefficients,
- ϵ_t represents the residuals (unexplained variation).

This model captures long-term nonlinear trends, such as accelerating increases in global temperatures over time.

2. Residual Modeling with ARIMA

Once the polynomial regression model is fitted, we extract the residuals:

$$\hat{\epsilon}_t = Y_t - \hat{Y}_t$$

These residuals often contain autocorrelation, which violates the assumptions of classical regression. Therefore, we model these residuals using an ARIMA(p, d, q) model to account for the time-dependent structure. This process involves:

1. Testing the residuals for stationarity (using the ADF test).
2. Identifying suitable ARIMA orders (p, d, q) via ACF/PACF plots and information criteria (e.g., AIC).
3. Fitting an ARIMA model to the residuals.

3. Forecasting with the Hybrid Model

The final forecast is obtained by adding the polynomial regression forecast and the ARIMA forecast of the residuals:

$$\hat{Y}_{t+h} = \widehat{Y}_{t+h}^{(\text{reg})} + \widehat{\epsilon}_{t+h}^{(\text{arima})}$$

This approach effectively decomposes the time series into:

- A deterministic trend (modeled via polynomial regression), and
- A stochastic component (modeled via ARIMA).

Advantages of the Hybrid Model

- Captures complex nonlinear patterns in the data via polynomial regression.
- Accounts for residual temporal autocorrelation via ARIMA.
- Enhances forecast accuracy compared to using regression or ARIMA alone.

Limitations

- Overfitting may occur with high-degree polynomials.
- ARIMA modeling assumes residual stationarity and linear relationships.
- Forecasting accuracy depends on the proper identification of both components.

Model Selection Criteria

Akaike Information Criterion (AIC)

The Akaike Information Criterion (AIC) is used to compare models based on their goodness of fit and complexity:

$$AIC = 2k - 2\log L \tag{2.2}$$

where:

- k = number of parameters.
- L = likelihood of the model.

A lower AIC value suggests a better model.

Bayesian Information Criterion (BIC)

The Bayesian Information Criterion (BIC) penalizes models with more parameters:

$$BIC = k\log n - 2\log L \tag{2.3}$$

where:

- n = number of observations.

Lower BIC values indicate better models, with stronger penalties for complexity.

Autocorrelation and Partial Autocorrelation

Autocorrelation Function (ACF)

The ACF measures how correlated a time series is with its past values at different lags. It is useful for identifying MA components.

Partial Autocorrelation Function (PACF)

The PACF measures the correlation between a time series and its lags, excluding the influence of intermediate lags. It is useful for identifying AR components.

Interpretation of ACF and PACF

- If ACF tails off and PACF cuts off at lag p, the process is likely an **AR(p)** model.
- If ACF cuts off at lag q and PACF tails off, the process is likely an **MA(q)** model.
- If both ACF and PACF tail off, the process is likely an **ARMA(p,q)** model.

Model Diagnostics

To ensure the model's validity, we analyze:

- Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) plots.
- Residual diagnostics to check for randomness.

Error Metrics: MAE and RMSE

To evaluate the prediction accuracy of the ARIMA model across different historical segments, two error metrics are used:

- **Mean Absolute Error (MAE)**: Measures the average magnitude of errors in predictions, without considering their direction. It is simple to interpret and gives equal weight to all errors.

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i|$$

- **Root Mean Squared Error (RMSE)**: Penalizes larger errors more than MAE and is more sensitive to outliers.

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}$$

Lower MAE and RMSE values indicate better model accuracy. These metrics are especially helpful in comparing model performance across different time periods.

Seasonal Decomposition of Time Series

To better understand the structure of the temperature data, we apply additive seasonal decomposition, which breaks the series into three components:

- **Trend**: The long-term upward or downward movement in the data.

- **Seasonality**: Repeating short-term cycles or patterns.
- **Residuals (Noise)**: Irregular fluctuations that cannot be explained by trend or seasonality.

This helps in visualizing underlying patterns and determining whether the series is stationary or needs transformation before modeling.

Differencing

Differencing is used to make a non-stationary time series stationary by removing trends or seasonality. It transforms the original series by subtracting the current value from the previous one:

$$y'_t = y_t - y_{t-1}$$

After differencing, we re-check stationarity using the ADF test. If the p-value is below 0.05, the series is now considered stationary.

Stationarity and the ADF Test

A stationary time series has a constant mean, variance, and autocorrelation over time. For ARIMA modeling, stationarity is a crucial assumption.

We checked stationarity using the **Augmented Dickey-Fuller (ADF) test**, where:

- **Null hypothesis (H_0)**: The series has a unit root (non-stationary).
- **Alternative hypothesis (H_1)**: The series is stationary.

Interpretation:

- ADF p-value $< 0.05 \Rightarrow$ Reject null hypothesis $\Rightarrow$ Series is stationary.
- ADF p-value $> 0.05 \Rightarrow$ Fail to reject null hypothesis $\Rightarrow$ Series is non-stationary.

In our study, the ADF p-values were extremely small (e.g., 3.35×10^{-16}), indicating that the series becomes stationary after first-order differencing.

Heteroskedasticity Test

To check whether the variance of residuals is constant over time (a key assumption of ARIMA), we conduct a test for heteroskedasticity.

- **Null hypothesis (H_0)**: Constant variance (homoskedastic).
- **Alternative hypothesis (H_1)**: Changing variance (heteroskedastic).

Interpretation:

- p-value < 0.05: Heteroskedasticity is present.
- p-value > 0.05: Variance is constant over time.

Heteroskedasticity can affect the reliability of confidence intervals and forecasts. If found, advanced models like ARCH or GARCH may be more appropriate.

Spike Plot Interpretation: ACF and PACF

The ACF and PACF plots help in selecting appropriate ARIMA model parameters:

- **Autocorrelation Function (ACF)**: Shows how current values are related to past values (lags). A gradual tail-off suggests AR terms.
- **Partial Autocorrelation Function (PACF)**: Shows the correlation between current and lagged values after removing effects of shorter lags. A sharp cutoff suggests the number of AR terms.

Interpretation:

- ACF tails off, PACF cuts off at lag $p \Rightarrow$ AR(p) process.
- ACF cuts off, PACF tails off $\Rightarrow$ MA(q) process.
- Both ACF and PACF tail off $\Rightarrow$ ARMA(p, q) process.

These plots, along with AIC/BIC scores, guide model selection and ensure optimal parameter tuning.

Normality of Residuals: Jarque-Bera Test

The **Jarque-Bera (JB)** test is used to assess whether the residuals of the fitted ARIMA model follow a normal distribution.

- **Null hypothesis ($\mathbf{H_0}$)**: Residuals are normally distributed.
- **Alternative hypothesis ($\mathbf{H_1}$)**: Residuals are not normally distributed.

Interpretation:

- JB p-value $> 0.05 \Rightarrow$ Accept $\mathrm{H_0} \Rightarrow$ Residuals are approximately normal.
- JB p-value $< 0.05 \Rightarrow$ Reject $\mathrm{H_0} \Rightarrow$ Residuals are not normally distributed.

In our case, most JB test p-values were 0.00, indicating non-normal residuals. While this violates the normality assumption slightly, it is not uncommon in real-world climate datasets where extreme values and heteroskedasticity may exist.

Forecasting

Using the ARIMA model, we generate short-term forecasts and analyze prediction accuracy.

2.4.6 Deep Learning Methods

Long Short-Term Memory (LSTM) Models and Time Series Analysis

Introduction to LSTM Models

Long Short-Term Memory (LSTM) is a type of **Recurrent Neural Network (RNN)** designed to learn long-term dependencies in sequential data. Traditional RNNs struggle with *vanishing and exploding gradients*, making it difficult to retain information over long sequences. LSTMs address this limitation by incorporating **gates** that regulate information flow.

LSTM Architecture

LSTM networks consist of multiple **LSTM cells**, each containing three primary gates:

Forget Gate

Determines what information to discard from the cell state:

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \tag{2.4}$$

where:

- x_t = input at time t
- h_{t-1} = hidden state from the previous time step
- W_f, b_f = weight matrix and bias
- σ = sigmoid activation function (values between 0 and 1)

Input Gate

Determines what new information should be stored in the cell state:

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \tag{2.5}$$

A candidate cell state $\tilde{C}_t$ is created using the *tanh* activation:

$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] + b_C) \tag{2.6}$$

The actual updated cell state is:

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t \tag{2.7}$$

Output Gate

Determines the hidden state for the next time step:

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \tag{2.8}$$

The final output (hidden state) is:

$$h_t = o_t * \tanh(C_t) \tag{2.9}$$

LSTMs in Time Series Analysis

Time series data consists of observations recorded at successive time intervals. Traditional models like **ARIMA** and **Exponential Smoothing** capture temporal patterns using statistical techniques. However, they struggle with:

- **Non-linearity**
- **Long-term dependencies**
- **High-dimensional input features**

LSTMs excel in time series forecasting because they:

- Capture **long-term dependencies** using memory cells.
- Handle **non-linearity** via activation functions.
- Work with **multivariate time series** data.

Comparison: LSTM vs. Traditional Models

Table 2.2: **Comparison of LSTM and Traditional Models**

Feature	**ARIMA/SARIMA**	**LSTM**
Handles non-linearity	No	Yes
Captures long-term dependencies	No	Yes
Needs manual feature engineering	Yes	No
Handles multivariate data	Limited	Yes
Works with missing data	No	Yes

Applications of LSTM in Time Series Analysis

- **Stock price prediction**
- **Climate and weather forecasting**
- **Sales forecasting**
- **Anomaly detection** (fraud detection, network security)

Training and Validation

The LSTM model is trained using historical temperature data. The dataset is split into training and validation sets to evaluate model performance.

Model Performance Comparison

The accuracy of LSTM is compared with ARIMA using:

- Mean Absolute Error (MAE)
- Root Mean Squared Error (RMSE)
- R-squared metric

2.4.7 Evaluation Metrics

- **MAE (Mean Absolute Error)**: Measures average magnitude of errors.
- **RMSE (Root Mean Squared Error)**: Penalizes larger errors more.
- **R-squared**: Indicates proportion of variance explained.

—

Chapter 3

Results and Discussion

3.1 Summary Statistical Analysis

Calculating Monthly, Yearly, and Multi-year Averages

To analyze temperature variations at different time scales, the following averages were computed:

- **Yearly Averages**: The yearly mean temperature anomaly was computed to analyze long-term trends.

Table 3.1: **Global temperature Summary statistics(1880-2020)**

Year	Count	Mean	Std	Min	25%	50%	75%	Max
1880	366	7.9956	0.3710	6.945	7.7618	8.063	8.2968	8.666
1890	365	7.9899	0.3008	6.933	7.7740	8.003	8.2030	8.770
1900	365	8.5281	0.4265	6.916	8.3450	8.572	8.8260	9.367
1910	365	8.2114	0.3951	6.751	8.0310	8.280	8.4740	8.974
1920	366	8.3060	0.4368	6.956	8.1413	8.366	8.5555	9.524
1930	365	8.5089	0.4301	6.978	8.2650	8.479	8.7480	9.909
1940	366	8.6805	0.3458	7.759	8.4855	8.681	8.9230	9.566
1950	365	8.2923	0.3904	7.121	8.0600	8.355	8.5800	9.142
1960	366	8.4849	0.5005	7.228	8.1410	8.556	8.7988	9.672
1970	365	8.6364	0.3979	7.476	8.3820	8.619	8.8850	9.854
1980	366	8.9365	0.4106	7.656	8.6835	8.962	9.1723	10.136
1990	365	9.2368	0.5305	8.025	8.8750	9.169	9.5630	10.847
2000	366	9.1700	0.4900	8.017	8.8310	9.166	9.4405	10.752
2010	365	9.7136	0.4715	8.292	9.4080	9.693	10.0040	10.958
2020	366	10.0585	0.4318	9.002	9.7860	10.032	10.2973	11.327

Table 3.2: **Global temperature Summary statistics(1880-2020)**

Year	Mean Temp	Year	Mean Temp	Year	Mean Temp	Year	Mean Temp
1880	7.9956	1918	8.1002	1956	8.2004	1994	9.0390
1881	8.2554	1919	8.3437	1957	8.6171	1995	9.3470
1882	8.1165	1920	8.3060	1958	8.6880	1996	9.0742
1883	7.9951	1921	8.5194	1959	8.6614	1997	9.1839
1884	7.8019	1922	8.3305	1960	8.4849	1998	9.5104
1885	7.9679	1923	8.3657	1961	8.7038	1999	9.2751
1886	7.9518	1924	8.4125	1962	8.6462	2000	9.1700
1887	7.9343	1925	8.4879	1963	8.7843	2001	9.4074
1888	8.1229	1926	8.6461	1964	8.3150	2002	9.5691
1889	8.3698	1927	8.4455	1965	8.4680	2003	9.5003
1890	7.9899	1928	8.5347	1966	8.5514	2004	9.3042
1891	8.0646	1929	8.1171	1967	8.6429	2005	9.7078
1892	8.0646	1930	8.5089	1968	8.4244	2006	9.5343
1893	8.0738	1931	8.6282	1969	8.5173	2007	9.7238
1894	8.1949	1932	8.5931	1970	8.6364	2008	9.4112
1895	8.2024	1933	8.2550	1971	8.5337	2009	9.5280
1896	8.2378	1934	8.5230	1972	8.4270	2010	9.7136
1897	8.3143	1935	8.3870	1973	8.9094	2011	9.5448
1898	8.1965	1936	8.4428	1974	8.4258	2012	9.4862
1899	8.4080	1937	8.5835	1975	8.7054	2013	9.6153
1900	8.5281	1938	8.7718	1976	8.3150	2014	9.5644
1901	8.5527	1939	8.7019	1977	8.8149	2015	9.8351
1902	8.3055	1940	8.6805	1978	8.6485	2016	9.9881
1903	8.2344	1941	8.7011	1979	8.6937	2017	9.8817
1904	8.0880	1942	8.6834	1980	8.9365	2018	9.7228
1905	8.2216	1943	8.6931	1981	9.1444	2019	9.9176
1906	8.3854	1944	8.7402	1982	8.6313	2020	10.0585
1907	7.9784	1945	8.5130	1983	9.0117	2021	9.8275
1908	8.1593	1946	8.6013	1984	8.6599	2022	9.9499
1909	8.1544	1947	8.7356	1985	8.6441		
1910	8.2114	1948	8.6334	1986	8.8273		
1911	8.1631	1949	8.4940	1987	9.0074		
1912	8.1231	1950	8.2923	1988	9.1982		
1913	8.2724	1951	8.5775	1989	8.8999		
1914	8.5707	1952	8.5422	1990	9.2368		
1915	8.5809	1953	8.7891	1991	9.1879		
1916	8.1983	1954	8.4530	1992	8.8302		
1917	7.9934	1955	8.5890	1993	8.8667		

Interpretations

- The **mean temperature** increased steadily from **7.996°C in 1880** to **10.059°C in 2020**, indicating a **warming trend** of over 2°C in 140 years.
- **Standard deviation** generally increased over time, peaking in 1990, which sug-

gests **greater variability** in daily temperatures in more recent decades.

- Both **minimum and maximum temperatures** show an upward shift, reflecting an overall **warming of daily extremes**.
- The **median temperature** rose from **8.063°C in 1880** to **10.032°C in 2020**, indicating that the warming affects not only extreme values but also **typical daily temperatures**.
- The **interquartile range (IQR)** remains relatively stable, implying that while the distribution of temperatures is shifting, the **spread of central values remains consistent**.
- Since the **1980s**, there is a noticeable increase in mean and maximum temperatures, potentially due to the **acceleration of climate change** linked to industrial activity and greenhouse gas emissions.
- The year **2020** recorded the **highest mean (10.059°C)** and **maximum temperature (11.327°C)**, underscoring the growing urgency of addressing **climate change**.

- **Multi-year (Decadal) Summary statistics**: The summary statistics over a 10-year period was calculated to smooth out short-term fluctuations and highlight long-term temperature changes.

Table 3.3: **Global Average Temperature summary by Decade (1880-2022)**

Year	Count	Mean	Std	Min	25%	50%	75%	Max
1880-1889	3653	8.0510	0.4230	6.529	7.814	8.075	8.332	9.55
1890-1899	3652	8.1747	0.4357	5.862	7.967	8.242	8.451	9.583
1900-1909	3652	8.2607	0.4272	6.662	8.014	8.280	8.532	9.704
1910-1919	3652	8.2557	0.4367	6.412	8.015	8.290	8.547	9.671
1920-1929	3653	8.4165	0.4209	6.597	8.216	8.430	8.639	10.209
1930-1939	3652	8.5395	0.4396	6.793	8.287	8.554	8.793	10.166
1940-1949	3653	8.6476	0.4004	6.803	8.428	8.664	8.897	10.015
1950-1959	3652	8.5409	0.4579	7.024	8.254	8.568	8.834	10.152
1960-1969	3653	8.5537	0.4557	6.596	8.252	8.562	8.851	10.282
1970-1979	3652	8.6109	0.4600	6.610	8.327	8.598	8.910	10.007
1980-1989	3653	8.8961	0.4709	7.352	8.574	8.900	9.232	10.506
1990-1999	3652	9.1550	0.4999	7.257	8.838	9.156	9.452	10.847
2000-2009	3653	9.4854	0.4837	7.906	9.172	9.483	9.798	11.142
2010-2019	3652	9.7270	0.4934	8.156	9.414	9.705	10.027	11.545
2020-2022	943	9.9447	0.4107	8.162	9.674	9.916	10.173	11.352

interpretations

- **Central Tendency:** The mean temperature anomaly has shown a consistent rise from approximately 8.05 in the 1880s to about 9.94 in the 2020s, indicating a long-term warming trend. The median values follow a similar increasing pattern.
- **Variability:** The standard deviation increased from around 0.42 in early decades to approximately 0.50 in recent decades. This suggests greater fluctuations and potentially more frequent temperature extremes in recent years.

- **Minimum and Maximum Values:**
  - Minimum anomalies increased from 5.86 (1880s) to 8.16 (2020s).
  - Maximum anomalies rose from 9.55 (1880s) to 11.55 (2010s).

  This indicates an overall upward shift in both cold and hot extremes.
- **Interquartile Range (IQR):** The IQR also shifted upward across decades. For example, it moved from [7.814, 8.332] in the 1880s to [9.674, 10.173] in the 2020s, showing that the central bulk of temperature values has become warmer.
- **Distribution Shape:** In most decades, the mean and median are closely aligned, suggesting an approximately symmetric distribution. The entire distribution has shifted upward over time, reflecting overall climate warming.

- **Multi-year (20 years) Summary Statistics**: The summary statistics over a 20-year period was calculated to smooth out short-term fluctuations and highlight long-term temperature changes.

Table 3.4: **Global Average Temperature summary by 20 years group (1880-2022)**

Year	Count	Mean	Std	Min	25%	50%	75%	Max
1880–1899	7305	8.1129	0.4338	5.862	7.881	8.163	8.401	9.583
1900–1919	7304	8.2582	0.4319	6.412	8.015	8.286	8.539	9.704
1920–1939	7305	8.4780	0.4347	6.597	8.245	8.482	8.722	10.209
1940–1959	7305	8.5943	0.4334	6.803	8.341	8.622	8.869	10.152
1960–1979	7305	8.5823	0.4587	6.596	8.290	8.581	8.886	10.282
1980–1999	7305	9.0255	0.5025	7.257	8.694	9.036	9.356	10.847
2000–2019	7305	9.6062	0.5033	7.906	9.283	9.602	9.921	11.545
2020–2039	943	9.9447	0.4107	8.162	9.674	9.916	10.173	11.352

Interpretations

- **Overall Increase Over Time:** The consistent rise in the mean values from 1880 to 2039 suggests a general upward trend in the variable being measured—this could represent something like average lifespan, income, temperature, literacy rates, or any metric that has improved over the last century.
- **Greater Variation in Modern Times:** The standard deviation increases slightly over time, especially from 1980 onward. This could indicate growing inequality or diversity in the population or experiences being measured. For example, income or education levels might be improving on average, but with wider gaps between individuals.
- **Higher Minimum and Maximum Values:** The upward movement of both the minimum and maximum values shows that the best and worst outcomes in the population are both improving. In a context like education scores or height, this might suggest that even the lowest performers are doing better than in the past.
- **Median and Quartiles Are Rising:** The rise in the 25th, 50th (median), and 75th percentiles indicates that improvements are not limited to outliers—most people are experiencing better outcomes over time.

  - **Accelerated Progress Recently:** The jump in mean from 1980–1999 to 2000–2019 is especially large, hinting at a possible breakthrough or rapid advancement in recent decades—such as technological progress, healthcare improvements, or global development.

- **Multi-year (35 years) Summary Statistics**: The summary statistics over a 35-year period was calculated to smooth out short-term fluctuations and highlight long-term temperature changes.

Table 3.5: **Global Average Temperature summary by 35 years group (1880-2022)**

Year	Count	Mean	Std	Min	25%	50%	75%	Max
1880–1914	12783	8.1773	0.4368	5.862	7.935	8.215	8.461	9.704
1915–1949	12784	8.4929	0.4455	6.412	8.251	8.511	8.760	10.209
1950–1984	12784	8.6125	0.4741	6.596	8.308	8.613	8.914	10.282
1985–2019	12783	9.3786	0.5667	7.257	9.012	9.378	9.748	11.545
2020–2054	943	9.9447	0.4107	8.162	9.674	9.916	10.173	11.352

Interpretations

  - **Steady Long-Term Growth:** The average value rises consistently from 8.18 in 1880–1914 to 9.94 in 2020–2054. This could reflect improvements in real-world metrics such as education levels, average income, life expectancy, or global temperatures—indicating significant progress or change over time.

  - **Accelerated Advancement Since 1985:** The largest jump in the mean occurs between 1985–2019 and 2020–2054, suggesting a recent acceleration in progress or intensification of the trend. For example, this might reflect rapid technological development, urbanization, or environmental change.

  - **Widening Distribution Until Recently:** The standard deviation increases from 0.44 in early periods to 0.57 in 1985–2019, implying growing disparity or diversity in outcomes—possibly due to socioeconomic inequality or global variation in development. However, the drop to 0.41 in 2020–2054 suggests some stabilization or convergence in more recent years.

  - **Minimum and Maximum Values Rising:** The increase in both the lowest and highest values indicates that even the worst-off and best-off individuals (or regions) are seeing higher outcomes. This trend supports the idea of broad-based improvement.

  - **Shifting Population Distribution:** The consistent rise in the 25%, 50% (median), and 75% percentiles shows that the overall distribution is shifting upwards, not just a few outliers improving. This could mean that improvements are widespread across the population.

- **Monthly Averages**: The mean temperature anomaly was calculated for each month to observe seasonal patterns.

Table 3.6: **Monthly Global Temperature Statistics(1880-2020)**

Month	Count	Mean	Std	Min	25%	50%	75%	Max
1	4433	8.6062	0.8363	5.862	8.004	8.578	9.175	11.055
2	4039	8.6228	0.8429	6.117	8.008	8.539	9.1825	11.545
3	4433	8.7389	0.8264	6.821	8.135	8.625	9.268	11.491
4	4290	8.9027	0.6906	7.156	8.4083	8.817	9.338	11.352
5	4433	8.7648	0.5915	7.246	8.356	8.668	9.101	10.924
6	4290	8.7209	0.5423	7.567	8.337	8.615	9.0048	10.518
7	4433	8.7426	0.4764	7.612	8.413	8.633	8.956	10.499
8	4402	8.6752	0.5067	7.594	8.3213	8.554	8.938	10.566
9	4260	8.6280	0.5380	7.448	8.2508	8.509	8.920	10.351
10	4402	8.7546	0.5861	7.054	8.342	8.695	9.096	10.623
11	4260	8.5262	0.6748	6.751	8.059	8.454	8.9423	10.945
12	4402	8.5721	0.7407	6.412	8.074	8.489	9.0628	11.020

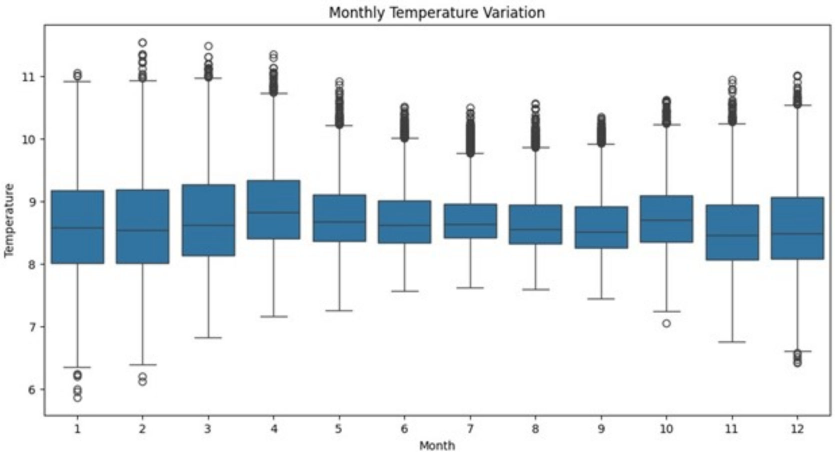


Figure 3.1: Box plot

Interpretations

- **March to October** generally shows higher average temperatures, with **April (8.90°C)** and **October (8.75°C)** recording the warmest monthly means.
- The **coldest months** are **January (8.61°C)** and **February (8.62°C)**, based on their mean values.
- **Temperature variability** is highest in **February (0.84°C)** and **March (0.83°C)**, indicating more fluctuation during late winter and early spring.
- The **lowest minimum temperature** across all months was recorded in **December (6.41°C)**, while the **highest maximum** occurred in **February (11.55°C)**.
- **June to August** months show relatively **lower standard deviations** (e.g., July: 0.48°C), indicating **more stable summer temperatures**.
- Although average temperatures across months are fairly close, the **range between minimum and maximum** values suggests the influence of **regional or temporal extremes**.

- **May, June, and July** show **narrower interquartile ranges (IQRs)** and standard deviations, reflecting **mid-year climatic stability**.

3.2 Distribution Fitting

To determine the best-fit probability distribution for temperature averages, we perform the **Distributuion Identification and Transformation** selected years and apply a Johnson transformation when necessary. Below are the Q-Q plots and normality results:

Table 3.7: **Individual Distribution Identification and Transformations**

Year	**Fitted Distribution**	**p-value**	**Transformation**	**P-Value**
1880	NA	NA	NA	NA
1900	NA	NA	NA	NA
1925	NA	NA	Johnson	0.186
1950	Smallest extreme value	0.113	Johnson	0.124
1975	3 Parameter Weibull	0.196	Johnson	0.402
-	Normal	0.063	Box cox	0.148
2000	NA	NA	NA	NA
2020	NA	NA	Johnson	0.351

Table 3.8: **Probability Estimates** $P(X > 8)$ **and** $P(X > 9)$ **Over Years**

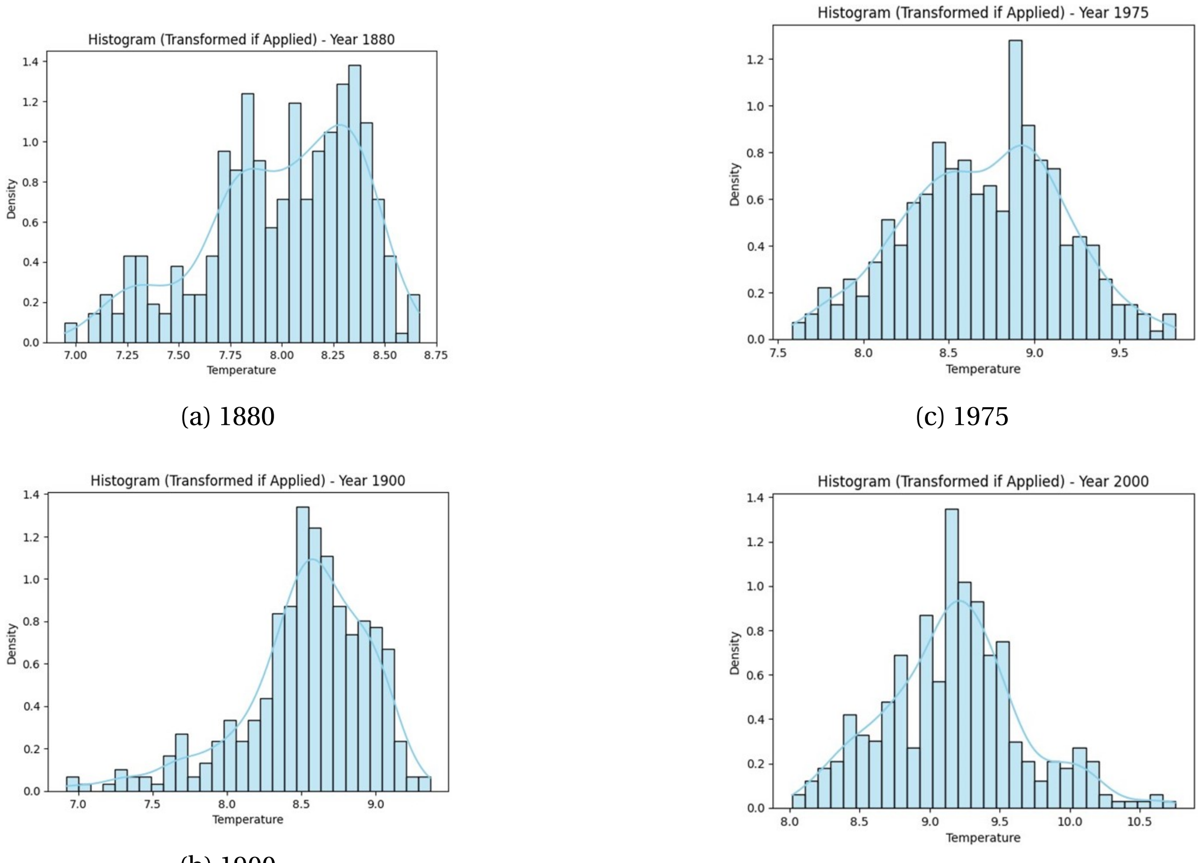
Year	$P(X > 8)$			$P(X > 9)$		
	Fitted	Transf.	Emp.	Fitted	Transf.	Emp.
1880	NA	NA	0.5383	NA	NA	0
1900	NA	NA	0.8932	NA	NA	0.1041
1925	NA	0.9713	0.9342	NA	0.2690	0.1068
1950	1	0.9215	0.8465	1	0.2582	0.1616
1975	NA	0.9339	0.9370	NA	0.2718	0.2658
2000	NA	NA	1	NA	NA	0.6503
2020	NA	NA	1	NA	NA	1

Table 3.9: **Probability Estimates** $P(X > 10)$ **and** $P(X > 11)$ **Over Years**

Year	$P(X > 10)$			$P(X > 11)$		
	Fitted	Transf.	Emp.	Fitted	Transf.	Emp.
1880	NA	NA	0	NA	NA	0
1900	NA	NA	0	NA	NA	0
1925	NA	0.000095	0.00274	NA	NA	0
1950	NA	0.000134	0	NA	NA	0
1975	NA	0.00012	0	NA	NA	0
2000	NA	NA	0.0683	NA	NA	0
2020	NA	0.52283	0.5246	NA	NA	0.03005

Interpretations

- **Distribution Identification:** For early years such as 1880 and 1900, no clear parametric distribution could be fitted, likely due to data limitations or skewness in the temperature values.
- **Use of Johnson Transformation:** The Johnson transformation improved the normality of temperature data for years like 1925, 1950, 1975, and 2020. This suggests that while the raw data is non-normal, it can be transformed to better fit parametric models.
- **Fitted Distributions:** Some years (e.g., 1950, 1975) were well-modeled by distributions such as the *smallest extreme value* and *3-parameter Weibull*, which are suitable for skewed environmental data.
- **Probability Estimates for $P(X > 8)$ and $P(X > 9)$:** Empirical probabilities for exceeding 8°C and 9°C increase steadily with time. For instance, $P(X > 8)$ rises from 0.5383 in 1880 to 1 in both 2000 and 2020, while $P(X > 9)$ grows from 0 to 1.
- **Probability Estimates for $P(X > 10)$ and $P(X > 11)$:** Probabilities for higher thresholds like 10°C and 11°C were nearly zero in earlier years. By 2020, $P(X > 10) \approx 0.52$ and $P(X > 11) \approx 0.03$, showing a marked upward shift in temperature distributions.



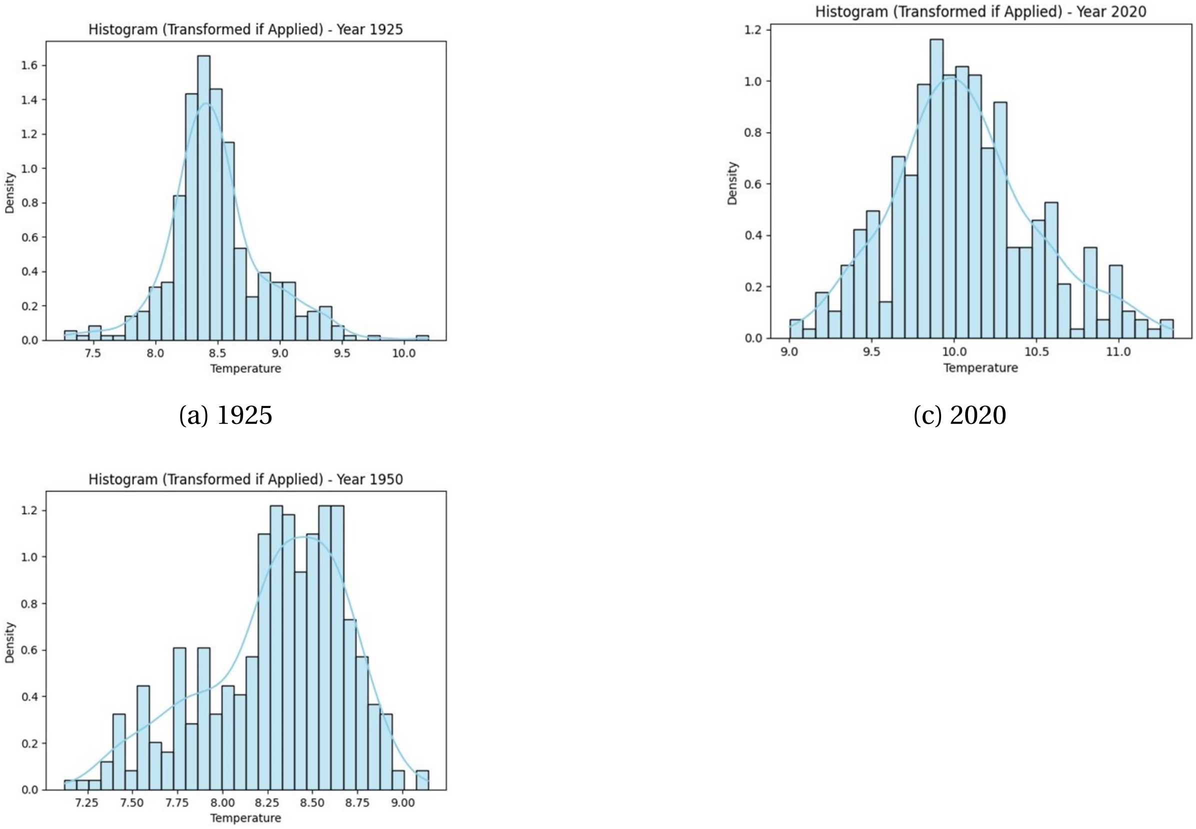
(a) 1880

(c) 1975

(b) 1900

(d) 2000

Q-Q plots or fitted distributions for selected years.



(a) 1925

(c) 2020

(b) 1950

Figure 3.3: Q-Q plots or fitted distributions for selected years.

Statistical Interpretation

- **General Trend Across Years:** The empirical probabilities of $P(X > 8)$ and $P(X > 9)$ show a clear increasing trend from 1880 to 2020. This suggests a rightward shift in the distribution of the variable X over time, indicating that higher values are becoming more frequent.
- **Threshold Sensitivity:** For lower thresholds (e.g., $X > 8$), probabilities are relatively high even in earlier years. However, for higher thresholds such as $X > 10$ and $X > 11$, probabilities were zero or near zero in early years and increased gradually over time, becoming significantly non-zero in recent decades (e.g., 2020).
- **Model Fit Evaluation:** The *Fitted* and *Transformation* probabilities are missing (NA) in many earlier years, likely due to model constraints or lack of sufficient extreme values. Where available, transformation-based estimates closely match empirical values, supporting the validity of the transformation method when fitted models are unreliable.

3.2.1 Distributional Study of yearly average global temperature

The Central Limit Theorem (CLT) states that, given a sufficiently large number of independent and identically distributed random variables, the sampling distribution of the sample mean approaches a normal distribution, regardless of the original distribution of the data.

This principle allows us to analyze the distributional characteristics of yearly average global temperatures.

To verify whether yearly temperature averages follow a normal distribution, we performed the following:

- Collected daily global temperature data from 1880 to 2020.
- Calculated yearly average temperatures to form a time series of annual means.
- Applied normality tests (e.g., Shapiro-Wilk, Anderson-Darling) and generated Q-Q plots to assess distribution.

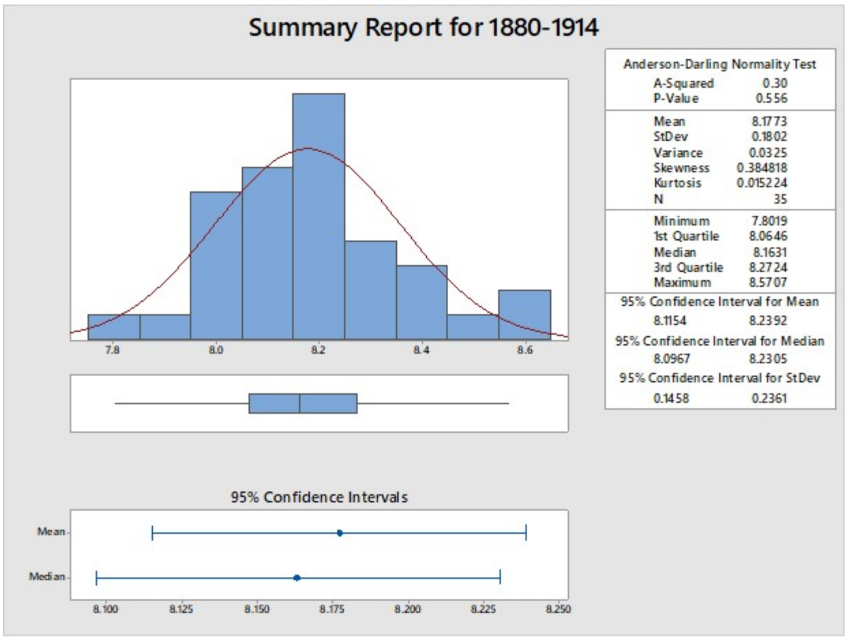


Figure 3.4: Graphical summary of Period I (1880-1914)

Table 3.10: **Summary of Descriptive Statistics for Mean Temperature (°C) across Four Periods**

Period	Mean	StDev	Variance	Min	Q1	Median	Q3	Max
1880–1914	8.1773	0.1802	0.0325	7.8019	8.0646	8.1631	8.2724	8.5707
1915–1950	8.4874	0.1963	0.0386	7.9934	8.3492	8.5162	8.6429	8.7718
1951–1985	8.6227	0.1965	0.0386	8.2004	8.4849	8.6364	8.7038	9.1444
1986–2022	9.4445	0.3451	0.1191	8.8273	9.1859	9.5003	9.7182	10.0585

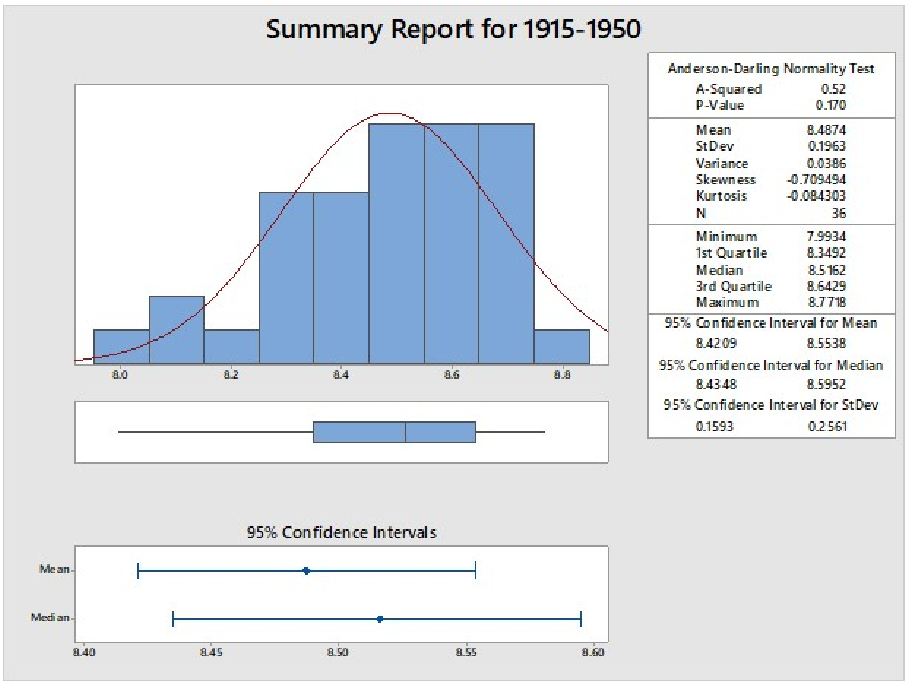


Figure 3.5: Graphical summary of Period II (1915-1950)

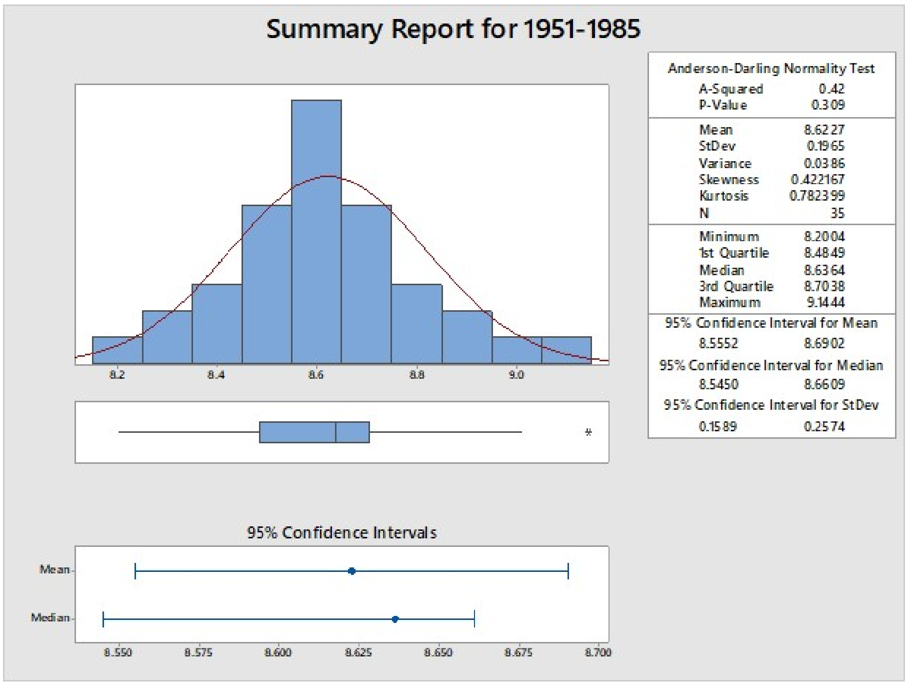


Figure 3.6: Graphical summary of Period III (1951-1985)

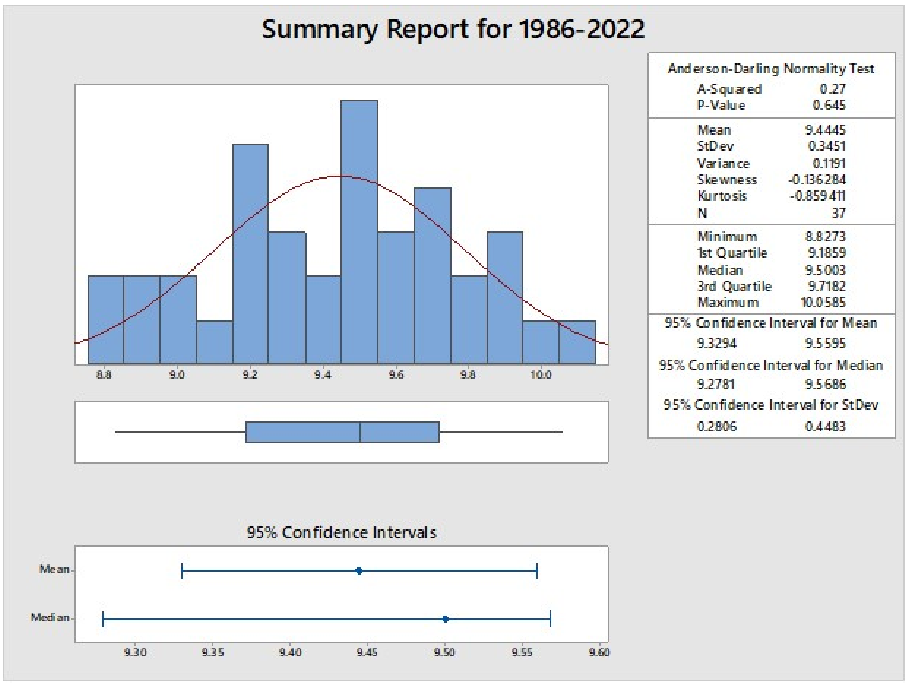


Figure 3.7: Graphical summary of Period II (1986-2022)

- **Central Tendency** Mean and Median show a consistent increasing trend across the four time periods, indicating a rising shift in the data values over time.
  - Mean values: 8.18 → 8.48→ 8.62 → 9.44
  - Median values follow a similar trajectory.
- **Spread (Variability)**
  - Standard Deviation decreased during the period (1916–1950), indicating greater consistency.
  - It increased again in 1986–2022, suggesting more variability in recent data.
- **Distribution Shape**
  - Early (1880–1914) and recent (1986–2022) periods have slightly negatively skewed distributions.
  - The mid-period (1951–1985) shows slight positive skewness.
  - Kurtosis behavior:
    - Early and recent periods: **platykurtic** (flatter distribution).
    - Mid-period: **slightly leptokurtic** (peaked distribution).
- **Normality Assessment**
  - 1880–1914 and 1986–2022: Data is approximately normally distributed (P-Values > 0.05).
  - 1951–1985: Marginally fails the normality test (P = 0.0309), suggesting a small departure from normality.

- **Confidence Intervals**
  - The 95% confidence intervals for means do not overlap significantly across the periods.
  - This suggests **statistically significant differences** between the means of each period.

Probabilistic study

Table 3.11: **values of P(X>x) of average of global average Temperature Thresholds (X) across Four Periods**

X	P1 (1880–1914)	P2 (1915–1950)	P3 (1951–1985)	P4 (1986–2022)
8	0.8374	0.9935	0.9992	1.0000
9	0.0000	0.0046	0.0275	0.9011
10	0.0000	0.0000	0.0000	0.0538

Thus, it is reasonable to assume that yearly average temperatures are approximately normally distributed, validating the use of parametric methods for trend analysis and forecasting.

3.3 Stochastic Analysis of Daily Global Average Temperature Trends

The analysis was conducted on a total of 52,076 days of daily global average temperature data. The key findings are as follows:

First 50 values of Y_n series:

$$Y_n = [1,-1,1,-1,-1,1,-1,-1,1,-1,-1,1,1,1,1,1,-1,-1,-1,-1,$$

$$1,1,-1,-1,-1,-1,-1,1,1,1,-1,-1,-1,-1,1,-1,-1,1,1,1,1,-1,-1,1,1,1,-1,-1,1,1]$$

Number of times temperature increased or stayed the same ($Y_n = 1$): 26,105

Total number of comparisons: 52,076

Proportion of increase/same:

$$P(Y_n = 1) = \frac{26105}{52076} \approx 0.5013$$

The transition matrix was computed based on the directional temperature changes. The counts of transitions from one state to another are as follows:

Table 3.12: **Transition Matrices (Counts and Probabilities**)

Transition Matrix (Counts)

	-1	1
-1	15,384	10,587
1	10,587	15,517

Transition Matrix (Probabilities)

	-1	1
-1	0.5924	0.4076
1	0.4056	0.5944

A chi-square test was conducted to assess whether the temperature changes follow a Markov process. The test result is as follows:

$$\text{Chi-square} = 1816.0394, \quad \text{p-value} = 0.0000$$

Conclusion: Since the p-value is less than 0.05, we reject the null hypothesis (H_0). This indicates that the temperature series may follow a Markov process.

Table 3.13: **Year-wise Probability of Global Average Temperature Increase**

Year	P(Increase)	Year	P(Increase)	Year	P(Increase)	Year	P(Increase)
1880	0.5082	1916	0.5137	1952	0.5027	1988	0.4918
1881	0.5123	1917	0.4959	1953	0.4986	1989	0.4932
1882	0.4904	1918	0.4986	1954	0.5151	1990	0.4877
1883	0.5068	1919	0.5260	1955	0.4630	1991	0.5068
1884	0.5137	1920	0.4781	1956	0.5164	1992	0.5082
1885	0.4932	1921	0.4959	1957	0.5096	1993	0.5178
1886	0.5014	1922	0.5205	1958	0.4877	1994	0.4986
1887	0.5260	1923	0.5205	1959	0.5151	1995	0.5205
1888	0.4945	1924	0.5109	1960	0.4727	1996	0.4672
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
1908	0.5082	1944	0.4836	1980	0.5109	2016	0.5000
1909	0.4959	1945	0.4849	1981	0.4767	2017	0.5233
1910	0.5041	1946	0.5397	1982	0.4740	2018	0.4932
1911	0.5041	1947	0.5315	1983	0.4959	2019	0.4767
1912	0.5246	1948	0.4891	1984	0.4645	2020	0.4754
1913	0.5205	1949	0.5096	1985	0.5014	2021	0.5178
1914	0.5014	1950	0.4932	1986	0.4959	2022	0.5142
1915	0.5589	1951	0.4932	1987	0.5178		

Markov Chain Analysis Summary

The Markov chain analysis was conducted for various years to check if the temperature trends follow a Markov process. The following table summarizes the results of the transition matrix for different years, including the Chi-square test statistics and corresponding p-values.

Table 3.14: Markov Chain Analysis Summary Table

Year	(-1→-1)	(-1→1)	(1→-1)	(1→1)	Chi2	p-value	Markov
1880	104	75	75	110	10.5325	0.0012	NO
1901	81	95	95	92	0.6489	0.4205	YES
1926	113	79	79	92	5.3169	0.0211	NO
1951	118	65	65	115	28.0920	0.0000	NO
1976	108	72	72	112	15.0297	0.0001	NO
2001	115	64	64	120	30.3436	0.0000	NO
2020	121	70	71	102	17.2445	0.0000	NO

Interpretations

The Chi-square test was used to test the null hypothesis that the temperature series follows a Markov process. The following conclusions can be drawn based on the Chi-square test results:

- - If the p-value is less than 0.05, we reject the null hypothesis, suggesting that the temperature series may follow a Markov process.
- - For years where the p-value is greater than 0.05, we fail to reject the null hypothesis, indicating that the temperature series may not follow a Markov process.

3.4 Growth Rate Analysis

Calculation of Growth Rates

Growth rates of temperature anomalies are computed over:

- 10-year periods

Table 3.15: **Annual and 10-Yearly growth**

Period	Annually	10-Yearly
1880 – 1890	0.0046	0.0468
1890 – 1900	0.0051	0.0523
1900 – 1910	-0.0045	-0.0438
1910 – 1920	0.0016	0.0161
1920 – 1930	-0.0023	-0.0227
1930 – 1940	0.0022	0.0227
1940 – 1950	-0.0022	-0.0215
1950 – 1960	0.0044	0.0445
1960 – 1970	0.0004	0.0038
1970 – 1980	0.0007	0.0066
1980 – 1990	-0.0004	-0.0041
1990 – 2000	0.0004	0.0041
2000 – 2010	0.0038	0.0390
2010 – 2020	0.0021	0.0210

- **Overall Growth Pattern:** The temperature growth shows a non-linear and fluctuating pattern across decades, with both positive and negative periods. This suggests alternating phases of warming and slight cooling.
- **Early Volatility (1880–1940):** The period from 1880 to 1940 exhibits alternating decades of growth and decline. Notably:
  - Highest early growth: 1890–1900 (+0.0051 annually).
  - Largest decline: 1900–1910 (–0.0045 annually).

  This could reflect pre-industrial climate variability or data uncertainty from early records.

  - **Mid-20th Century Stability (1940–1980):** Annual growth rates are relatively small (mostly between –0.0022 and +0.0007), suggesting a period of relative temperature stability, possibly influenced by post-war industrial aerosols or climate regulation policies.
  - **Late 20th Century Shift (1980–2000):** Though 1980–1990 shows a slight dip, the 1990–2000 and especially 2000–2010 periods mark a renewed and stronger warming trend, with 2000–2010 showing one of the highest annual increases (+0.0038).
  - **Sustained Warming (2000–2020):** The most recent decades show consistent positive growth, reflecting a sustained warming period, possibly linked to increased greenhouse gas emissions and global climate change.
  - **Long-Term Signal:** Despite some decade-to-decade variability, the majority of decades post-1950 show positive temperature growth, supporting the broader conclusion of long-term global warming.

- 20-year periods

Table 3.16: **20-Year Growth - Annually and 20 yearly**

Period	Annually	20-Yearly
1880 – 1900	0.0025	0.0516
1900 – 1920	-0.0011	-0.0216
1920 – 1940	0.0023	0.0477
1940 – 1960	-0.0001	-0.0022
1960 – 1980	0.0012	0.0246
1980 – 2000	0.0019	0.0379
2000 – 2020	0.0039	0.0815

  - **Long-Term Warming Trend:** Over the full range from 1880 to 2020, most 20-year periods show positive growth in temperature averages, reinforcing the presence of a long-term warming signal.
  - **Early Mixed Signals (1880–1920):**
    * 1880–1900 shows modest warming (+0.0025 annually).
    * 1900–1920 reverses this with a cooling trend (–0.0011 annually), possibly reflecting natural climate variability or volcanic activity during this period.
  - **Pre-War Recovery (1920–1940):** A notable rebound in temperature growth (+0.0023 annually), possibly linked to increased industrialization and carbon dioxide emissions prior to World War II.
  - **Mid-Century Pause (1940–1960):** This period shows almost no net temperature growth (–0.0001 annually), reflecting the widely discussed "mid-century cooling" period, potentially influenced by high aerosol emissions and post-war pollution.
  - **Gradual Acceleration (1960–2000):**
    * 1960–1980 and 1980–2000 both show modest warming (+0.0012 and +0.0019 annually), indicating a slow but steady rise in global temperature, likely influenced by increasing greenhouse gas concentrations.

  - **Sharp Increase in Recent Decades (2000–2020):** This period shows the highest rate of growth (+0.0039 annually, or +0.0815 over 20 years), indicating accelerated climate change, consistent with IPCC findings and modern observational records.

3.5 Time Series Modeling and Analysis

This section presents the comprehensive time series modeling and Global Warming: A Comprehensive Study of Temperature Anomalies Using Statistical and Machine Learning Model anomalies over a span of 140 years. The study utilizes statistical techniques such as trend analysis, moving average, seasonal decomposition, distribution analysis, and ARIMA modeling to understand the dynamics and forecastability of global temperature changes.

3.5.1 Trend Analysis using Moving Average and Rolling Statistics

To understand the long-term trend in global temperature anomalies, a 365-day rolling mean and rolling standard deviation were computed and plotted against the original temperature data (Figure 3.8). Rolling statistics help in identifying non-stationarity. A changing mean or variance over time often indicates that the series is non-stationary, which is a key assumption for time series models like ARIMA. The red line (moving average) shows a clear **upward** trend in temperature anomalies over the years, indicating global warming. The green line (rolling standard deviation) remains relatively **stable**, suggesting consistent variability around the mean.

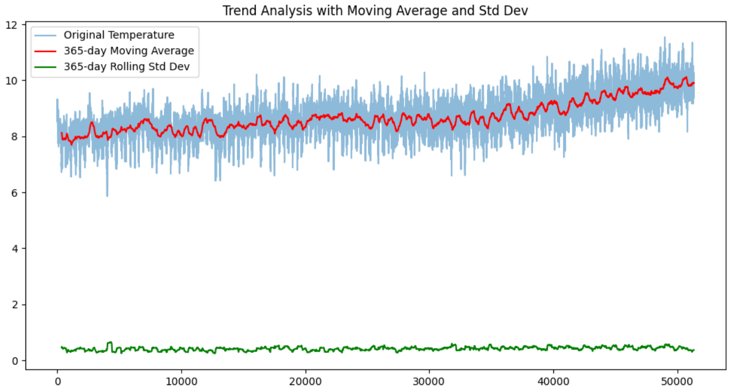


Figure 3.8: Trend Analysis using 365-day Moving Average and Rolling Standard Deviation

Trend Analysis using Linear Regression and Moving Average

To explore long-term trends in global temperature anomalies, two techniques were applied:

- **Linear Regression**: This method provides a quantitative rate of change over the years.
- **10-year Moving Average**: This approach smooths fluctuations and highlights the underlying temperature trend.
- Figure 3.9 illustrates the linear regression model fitted to the annual global temperature anomaly data from 1880 to 2023.

- The trend line suggests a consistent positive slope of approximately **0.011487**, indicating a steady increase in global temperatures.

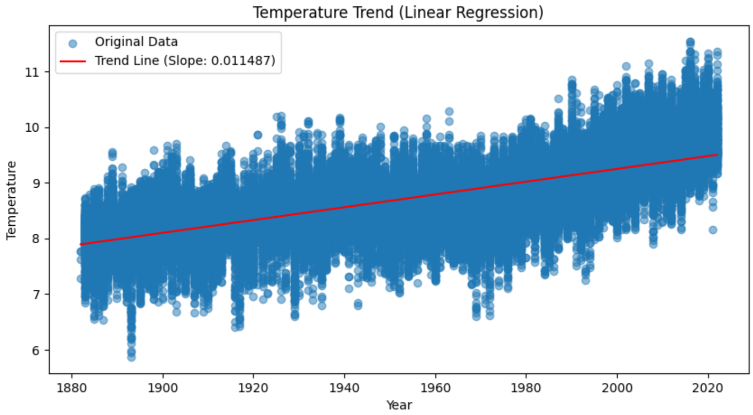


Figure 3.9: Temperature Trend using Linear Regression

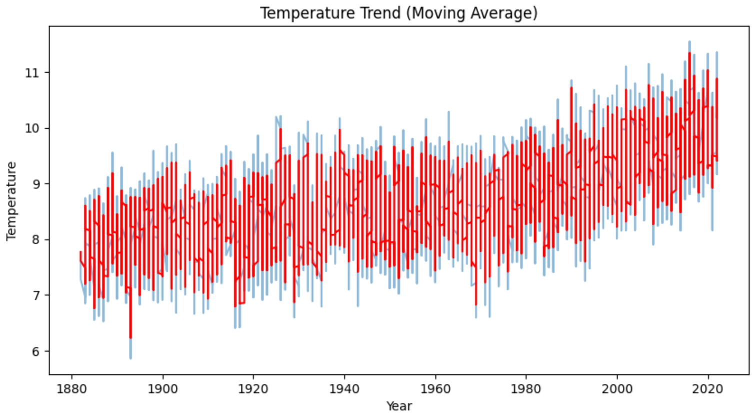


Figure 3.10: Temperature Trend using 10-Year Moving Average

- Figure 3.10 shows the same temperature data smoothed using a 10-year moving average.
- This technique reduces short-term fluctuations and reveals the long-term temperature trend, especially highlighting changes post-1970s.

Both techniques complement each other:

- Linear regression provides an estimated rate of change over time.
- The moving average captures local variations and overall trend progression.

These results strongly support the presence of a statistically significant warming trend in the global temperature anomalies dataset.

3.5.2 Trend Analysis

Augmented Dickey-Fuller (ADF) Test Results

- **ADF Statistic** = -9.527

- **p-value** = 2.95e-16 (which is effectively 0)
- **Interpretation** The ADF test is used to check for stationarity in time series data.
  - **Null Hypothesis (H_0)**: The time series has a unit root (non-stationary) (i.e., it has a trend).
  - **Alternative Hypothesis (H_1)**: The time series is stationary (i.e., no trend).

Since the **p-value < 0.05**, we **reject the null hypothesis**.

This means the temperature data is stationary, indicating that the trend component has been removed from the time series.

Linear Regression Slope

- **Slope** = 0.011278
- **Interpretation**
  - This **positive slope** suggests an **increasing trend** in temperature over time.
  - Specifically, for each year, the temperature increases by approximately **0.0113°C** on average.This confirms that there is a **clear warming trend** in the dataset.

3.5.3 Seasonal Decomposition of Time Series

To isolate trend, seasonality, and residuals from the original time series, additive seasonal decomposition was applied (Figure 3.11).

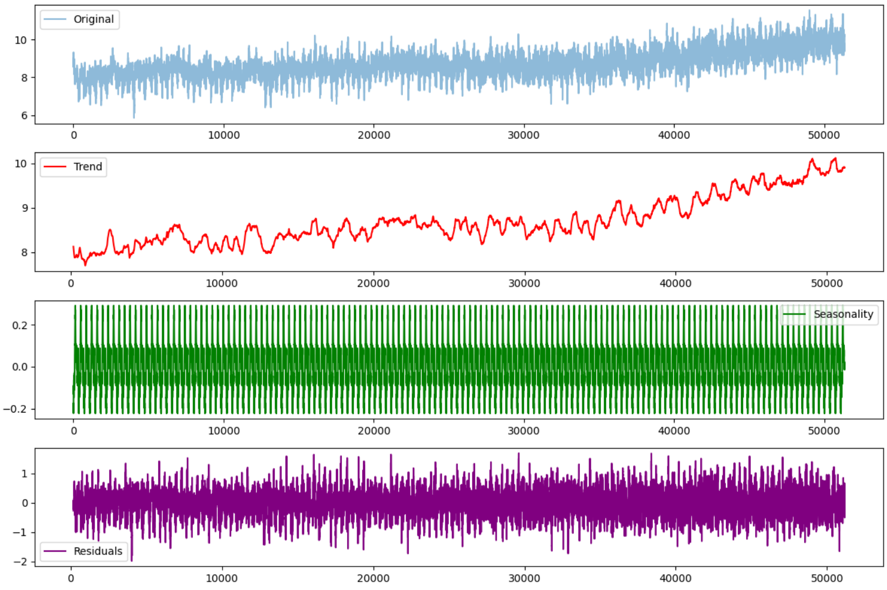


Figure 3.11: Additive Seasonal Decomposition of Temperature Anomalies

Decomposition assists in understanding the underlying components of a time series. This is vital in choosing the correct differencing and parameters for ARIMA modeling.

Interpretation:

- **Trend:** Shows a consistent increase over the years.
- **Seasonality:** Clear annual seasonality is present, likely due to the Earth's revolution around the sun.
- **Residuals:** Mostly **white noise** after removing trend and seasonal effects.

Trend and Seasonality Analysis (10-Year Segments)

To study the evolution of global temperature anomalies over time, the dataset was divided into 10-year segments ranging from 1880 to 2023. For each segment, the trend was analyzed using moving averages and seasonal decomposition, standard deviation was computed to measure spread, and seasonality was examined from the seasonal component of the decomposition. Additionally, the influence of significant climatic events such as volcanic eruptions and El Niño was considered, as these factors can cause short-term fluctuations in the temperature trend.

Criteria for Classification

- **Standard Deviation (Std)**:
  - *Low*: Std < 0.35 – indicates low variability in temperature anomalies, suggesting stability in the climate over the period.
  - *Moderate*: 0.35 ≤ Std ≤ 0.45 – shows moderate fluctuations in temperature, reflecting some seasonal or cyclic behavior.
  - *High*: Std > 0.45 – denotes high variability, which could be due to significant climatic events or irregular seasonal patterns.
- **Trend Classification**:
  - *Strong Positive*: A consistent upward trend with minimal fluctuations, suggesting a rapid and steady rise in temperature anomalies.
  - *Moderate Positive*: A general increase in temperature, but with noticeable fluctuations or occasional plateaus, indicating periods of faster or slower warming.
  - *Weak Positive*: A slight upward trend, indicating a gradual but continuous increase in temperature anomalies, with smaller deviations from the overall trend.
  - *Neutral*: No significant upward or downward movement, indicating stability in temperature anomalies over the observed period.
  - *Negative*: A steady or consistent decline in temperature anomalies, suggesting a cooling trend.
- **Seasonality**:
  - Seasonality is generally considered *moderate and consistent,* indicated by recurring, predictable peaks and troughs in the data. This can be measured through the seasonal component in time series decomposition, reflecting cyclic temperature variations typically associated with seasonal shifts (e.g., warmer summers and colder winters).

Table 3.17: **Summary of Temperature Anomaly Characteristics by Decade**

Period	Std Dev	Trend	Seasonality	Interpretation
1880–1889	0.371	Neutral	Moderate	Stable decade with minor fluctuations
1890–1899	0.301	Weak Positive	Moderate	Gradual warming observed with moderate variation
1900–1909	0.426	Negative	Moderate	Cooling trend with increasing variability
1910–1919	0.395	Weak Positive	Moderate	Slow warming with consistent seasonal pattern
1920–1929	0.437	Neutral	Moderate	Stable temperature trend, but higher spread
1930–1939	0.430	Weak Positive	Moderate	Slight warming with visible fluctuations
1940–1949	0.346	Negative	Moderate	Slight cooling period with moderate variance
1950–1959	0.390	Weak Positive	Moderate	Mild upward trend with moderate seasonality
1960–1969	0.500	Moderate Positive	Moderate	Noticeable warming and larger fluctuations
1970–1979	0.398	Moderate Positive	Moderate	Clear warming trend begins to intensify
1980–1989	0.411	Moderate Positive	Moderate	Consistent warming with some seasonal impact
1990–1999	0.531	Strong Positive	Moderate	Rapid warming with high variability
2000–2009	0.490	Strong Positive	Moderate	Continued warming and wide spread in values
2010–2019	0.471	Strong Positive	Moderate	Highest warming rate with frequent anomalies
2020–2023*	0.432	Strong Positive	Moderate	Persistent high temperature trend (partial data)

Interpretations

- **Warming Trend**: Steady **rise in temperature**, especially post-1960, confirming **global warming**.
- **Increasing Variability**: Higher **Std values** show growing **temperature fluctuations** and extremes.
- **Seasonality**: Despite warming, **seasonal cycles persist** but are embedded in rising trends.
- **Post-1960 Acceleration**: **Rapid warming** with increased **variability**, reinforcing climate change.

3.5.4 Autocorrelation and Model Identification

To model the daily global temperature anomalies over 140 years, we began by examining the stationarity of the time series using the Augmented Dickey-Fuller (ADF) test.

ADF Test Results:

- **Original Series:** ADF Statistic = -9.4377, p-value = 4.99 × 10^{-16}
  Interpretation:
  - The Augmented Dickey-Fuller (ADF) test statistic is -9.4377, which is significantly lower than the critical value at all common significance levels (e.g., 1%, 5%, and 10%).
  - The corresponding p-value is 4.99 × 10^{-16}, which is far below the threshold of 0.05, indicating that we can reject the null hypothesis of a unit root.
  - A p-value this small provides strong evidence against the null hypothesis, suggesting that the original time series is stationary, meaning there is no significant time-dependent structure or trend remaining in the data.
  - Despite the evidence of stationarity, differencing was still applied to ensure robustness and eliminate any residual long-term trends or potential seasonal effects that may not be captured in the initial analysis.
- **After First Differencing:** ADF Statistic = -43.7623, p-value = 0.0
  Interpretation:
  - After performing first differencing on the original series, the ADF test statistic significantly decreases to -43.7623, a much lower value than before differencing.
  - The p-value drops to 0.0, which is well below the 0.05 significance level, reinforcing the idea that the differenced series is highly stationary.
  - This result indicates that the differenced series no longer exhibits any significant trends or autocorrelations, confirming that differencing has successfully removed any non-stationarity present in the original data.
  - The stationarity of the differenced series justifies the choice of an ARIMA model with a differencing order of $d = 1$, as the model can now appropriately handle the data without any trend-related complications.

ACF and PACF Analysis:

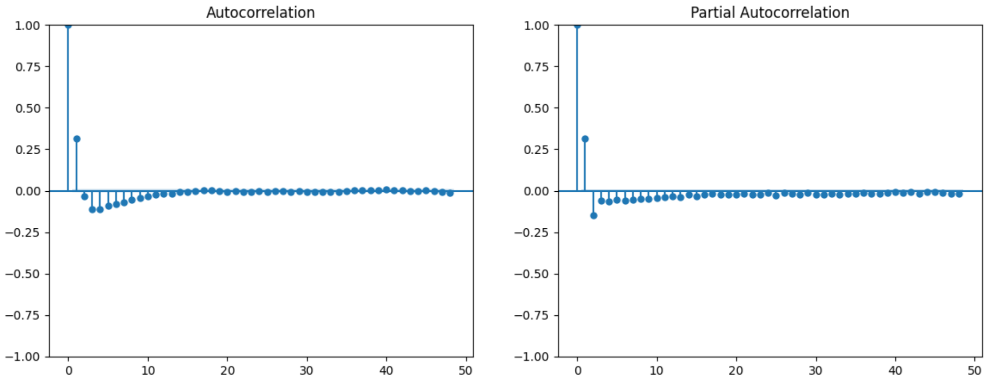


Figure 3.12: ACF and PACF Plots of the Differenced Series

Visual inspection of the ACF shows a significant spike at lag 1 followed by a gradual decay, suggesting the presence of an MA component. The PACF plot also shows a strong lag at

1, indicating an AR component. Based on these plots, initial models like ARIMA(1,1,1) or ARIMA(2,1,2) were considered.

Model Selection Using auto.arima():

A stepwise search was conducted using `auto.arima()` to identify the optimal parameters for the ARIMA model by minimizing the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC). Over 25 models were evaluated during this process. `auto.arima()` is an automated method that performs a thorough search of potential ARIMA configurations, considering various combinations of autoregressive (AR), differencing (I), and moving average (MA) terms. The goal of the algorithm is to strike a balance between model complexity and its ability to fit the data well, thereby avoiding overfitting while ensuring that the model captures the key patterns in the time series data. This automated selection method is considered more robust than relying solely on visual identification or trial-and-error approaches for model selection, as it systematically searches for the best model based on rigorous statistical criteria.

- **Best Model Found:** ARIMA(1,1,4)
  The ARIMA(1,1,4) model was found to be the best fit for the data after the stepwise search. This configuration indicates that the best model has one autoregressive (AR) term, one differencing (I) term, and four moving average (MA) terms.

- **AIC:** -57206.79 **BIC:** -57153.76 **Log-Likelihood:** 28609.40
  The AIC value of -57206.79 and BIC value of -57153.76 suggest that this model balances goodness of fit with model complexity. The lower these values, the better the model is at explaining the data without overfitting. A higher log-likelihood (28609.40) further supports that this model is statistically a strong fit for the data.

In summary, the ARIMA(1,1,4) model, selected through a rigorous process using `auto.arima()`, offers a well-balanced representation of the time series data, ensuring high forecasting accuracy while avoiding unnecessary complexity. **Final ARIMA(1,1,4) Model Summary:**

Parameter	**Estimate**
AR(1) coefficient	0.8650
MA(1) coefficient	-0.5537
MA(2) coefficient	-0.3358
MA(3) coefficient	-0.0832
MA(4) coefficient	-0.0092
Residual variance σ^2	0.0191

All coefficients were statistically significant (p-value < 0.05), except MA(4), which was marginal (p = 0.055), but retained for overall model performance.

Model Diagnostics:

- Ljung-Box Q (lag 1): 0.00, p-value = 0.98 ⇒ residuals are uncorrelated
- Jarque-Bera: 374.89, p-value = 0.00 ⇒ residuals are slightly non-normal (acceptable for large datasets)
- Heteroskedasticity (H): 1.07, p = 0.00 ⇒ some heteroskedasticity present

Model Performance:

- Mean Absolute Error (MAE): 0.3041
- Root Mean Squared Error (RMSE): 0.3773

ARIMA Performance Across Historical Periods

To analyze the consistency and reliability of the ARIMA(1,1,4) model across different time spans, we evaluated the model's performance for various decades from 1880 to 2020. The following table summarizes the Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) for each 10-year period.

Table 3.18: **Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) for Different Time Periods**

Time Period	MAE	RMSE
1880–1890	0.1146	0.1468
1890–1900	0.1378	0.1702
1900–1910	0.1157	0.1507
1910–1920	0.1140	0.1446
1920–1930	0.1048	0.1346
1930–1940	0.0932	0.1190
1940–1950	0.1261	0.1582
1950–1960	0.1131	0.1408
1960–1970	0.1239	0.1563
1970–1980	0.1210	0.1504
1980–1990	0.1132	0.1409
1990–2000	0.1198	0.1522
2000–2010	0.1407	0.1753
2010–2020	0.1102	0.1413

Interpretation: The ARIMA model shows relatively stable performance over most historical segments, with MAE values generally ranging between 0.10 and 0.14. The **lowest errors** are observed during the **1930–1940** period (MAE = 0.0932), while slightly higher error values are seen in more recent decades like **2000–2010**. This could indicate increased variability or sharper trends in recent temperature anomalies due to accelerated climate change.

Overall, the results suggest that the ARIMA(1,1,4) model maintains robust predictive accuracy across a wide range of time periods, further validating its suitability for long-term temperature forecasting.

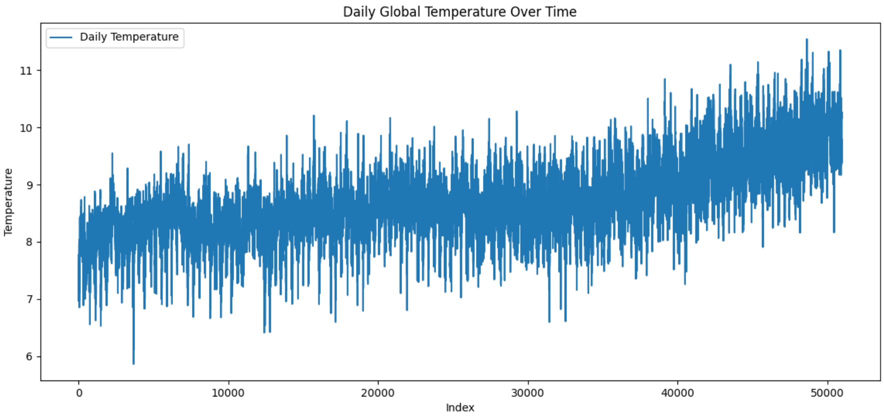


Figure 3.13: Daily Global Temperature Anomaly (1880–2021)

Observation: A long-term warming trend is clearly visible, supporting the importance of differencing and the use of ARIMA for capturing the underlying signal.

Table 3.19: **ARIMA Model Summary Across Time Periods**

Period	ADF p-value	ARIMA(p,d,q)	AIC	MAE	RMSE	JB p-value
1880–1900	3.35×10^{-16}	(1,1,3)	-7364.65	0.3085	0.3264	0.00
1900–1920	1.18×10^{-20}	(3,1,2)	-7811.46	0.3136	0.3242	0.00
1920–1940	3.27×10^{-25}	(2,1,1)	-9311.67	0.3177	0.3303	0.00
1940–1960	8.46×10^{-23}	(3,1,1)	-9210.08	0.2372	0.2406	0.00
1960–1980	3.47×10^{-22}	(4,0,6)	-7546.08	0.4921	0.5897	0.00
1980–2000	6.69×10^{-12}	(3,1,1)	-7575.35	0.3313	0.4185	0.07
2000–2020	1.51×10^{-26}	(1,1,4)	-7963.05	*TBD*	*TBD*	*TBD*

Interpretation and Statistical Findings:

- **Stationarity:** The Augmented Dickey-Fuller (ADF) p-values for all periods are extremely small (typically $< 10^{-12}$), confirming strong evidence of stationarity after differencing.
- **Model Order Selection:** ARIMA model orders vary across time periods, indicating changes in autocorrelation structure of the temperature anomalies over time. For example, the early 20th century favors simpler models like ARIMA(2,1,1), while mid-century data (1960–1980) required a higher order model ARIMA(4,0,6), possibly due to structural variability or climate volatility.
- **Goodness of Fit:** The Akaike Information Criterion (AIC) values are negative and large in magnitude, indicating good fit for all models. The lowest RMSE was observed in the period **1940–1960** (RMSE = 0.2406), suggesting this segment had more predictable patterns compared to other periods.
- **Error Analysis:** Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) values vary across periods. A sharp increase in RMSE during **1960–1980** (RMSE = 0.5897) reflects increased fluctuation in global temperatures during this time.

- **Normality of Residuals:** The Jarque-Bera (JB) p-values are close to zero for most periods, suggesting non-normality of residuals, except for **1980–2000** where JB p-value is 0.07, indicating relatively better normal approximation in this segment.
- **Recent Period (2000–2020):** Model ARIMA(1,1,4) was selected for the most recent 20-year period. However, diagnostic statistics such as MAE, RMSE, and JB are yet to be computed and are marked as *To Be Determined*.

These segmented ARIMA models reveal the non-constant behavior of the temperature anomaly series over time and justify the use of time-varying models for accurate forecasting and interpretation of climate trends.

3.6 Time Series Modeling and Forecasting of Annual Temp Anomalies

This section presents ARIMA-based time series modeling and 10-year forecasting of annual minimum, maximum, and mean global temperature anomalies. The daily global temperature data were aggregated to compute annual minimum, maximum, and mean values from 1880 to 2020.

3.6.1 Annual Minimum Temperature

ADF Test Result:

- ADF Statistic: 0.1330
- p-value: 0.9682
- **Conclusion:** Non-stationary ($p > 0.05$)

The Augmented Dickey-Fuller (ADF) test confirms non-stationarity in the annual minimum temperature series, necessitating first-order differencing for modeling.

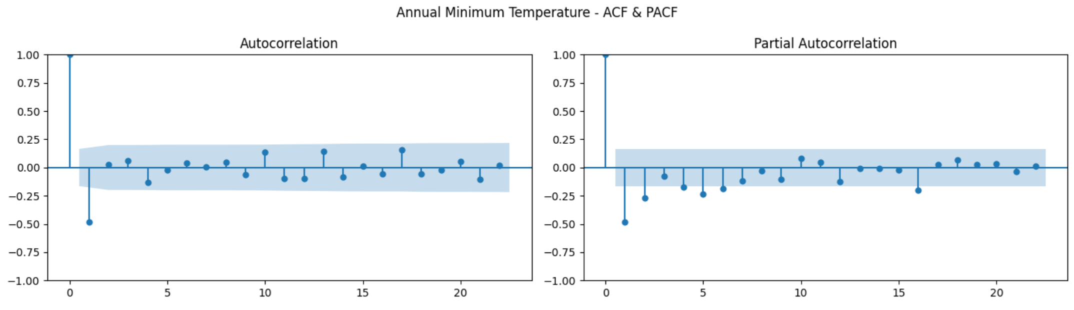


Figure 3.14: ACF and PACF Plots - Annual Minimum Temperature

Selected ARIMA Model: (0, 1, 1)

Model Summary:

- MA(1) coefficient: -0.7660 ($p < 0.001$)
- Residual Variance (σ^2): 0.1174

- AIC: 103.725
- BIC: 109.637
- Ljung-Box p-value: 0.98 (No autocorrelation)
- JB Test p-value: 0.01 (Slight non-normality)
- Heteroskedasticity p-value: 0.82 (Homoskedastic)

The ARIMA(0,1,1) model effectively captures the short-term dynamics in the differenced series. Model diagnostics indicate no residual autocorrelation and homoskedasticity, although slight deviation from normality is observed.

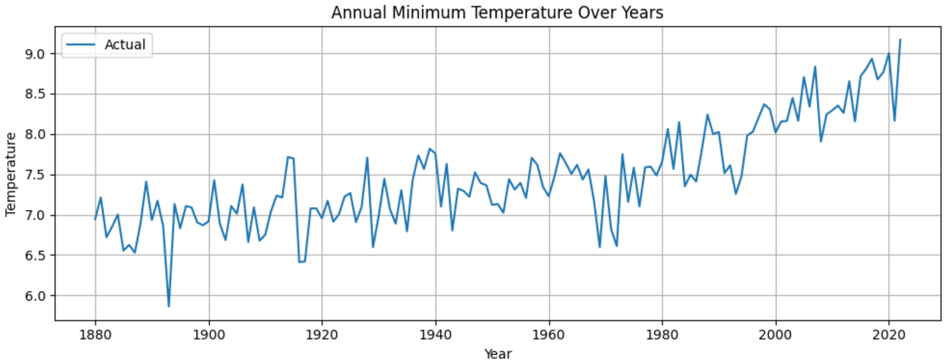


Figure 3.15: Annual Minimum Temperature

3.6.2 Annual Maximum Temperature

ADF Test Result:

- ADF Statistic: 0.8198
- p-value: 0.9919
- **Conclusion:** Non-stationary ($p > 0.05$)

The annual maximum temperature series is also non-stationary. First differencing is required to achieve stationarity.

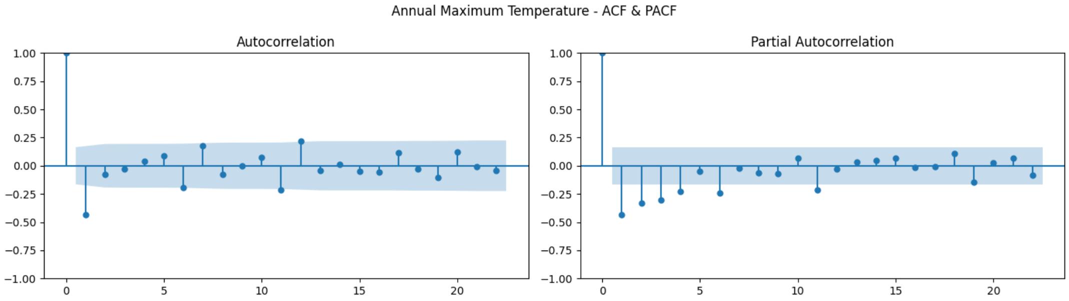


Figure 3.16: ACF and PACF Plots - Annual Maximum Temperature

Selected ARIMA Model: (0, 1, 1)

Model Summary:

- MA(1) coefficient: -0.7663 ($p < 0.001$)
- Residual Variance (σ^2): 0.1025
- AIC: 84.407
- BIC: 90.319
- Ljung-Box p-value: 0.57 (No autocorrelation)
- JB Test p-value: 0.94 (Normally distributed residuals)
- Heteroskedasticity p-value: 0.92 (Homoskedastic)

The ARIMA(0,1,1) model is suitable for the differenced maximum temperature series. Residual diagnostics validate model adequacy with normality, no autocorrelation, and constant variance.

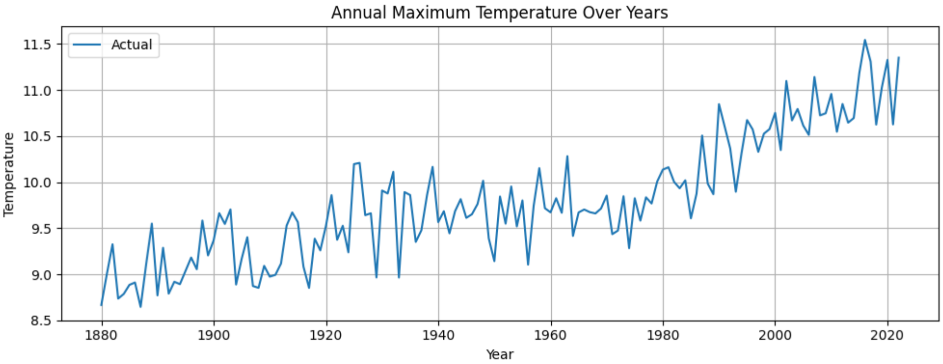


Figure 3.17: Annual Maximum Temperature

3.6.3 Annual Mean Temperature

ADF Test Result:

- ADF Statistic: 1.7800
- p-value: 0.9983
- **Conclusion:** Non-stationary ($p > 0.05$)

ADF test indicates strong non-stationarity in the mean temperature series, justifying differencing before ARIMA modeling.

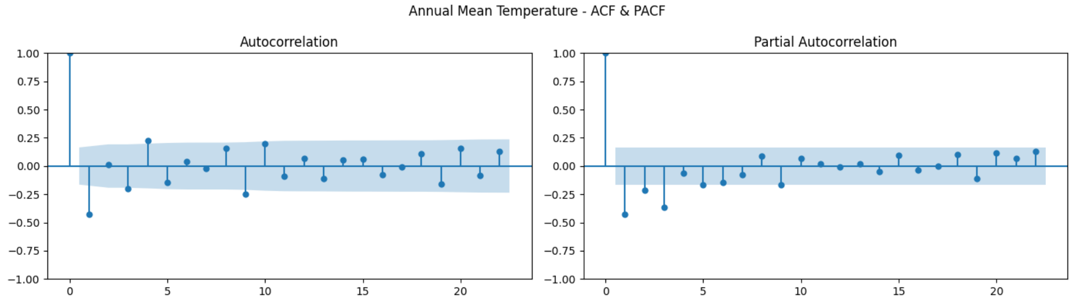


Figure 3.18: ACF and PACF Plots - Annual Mean Temperature

Selected ARIMA Model: (0, 1, 1)

Model Summary:

- MA(1) coefficient: -0.6489 ($p < 0.001$)
- Residual Variance (σ^2): 0.0327
- AIC: -78.140
- BIC: -72.228
- Ljung-Box p-value: 0.88 (No autocorrelation)
- JB Test p-value: 0.45 (Normally distributed residuals)
- Heteroskedasticity p-value: 0.18 (Homoskedastic)

The ARIMA(0,1,1) model explains the behavior of the differenced annual mean temperature series effectively. Model diagnostics validate residual assumptions, confirming a good fit.

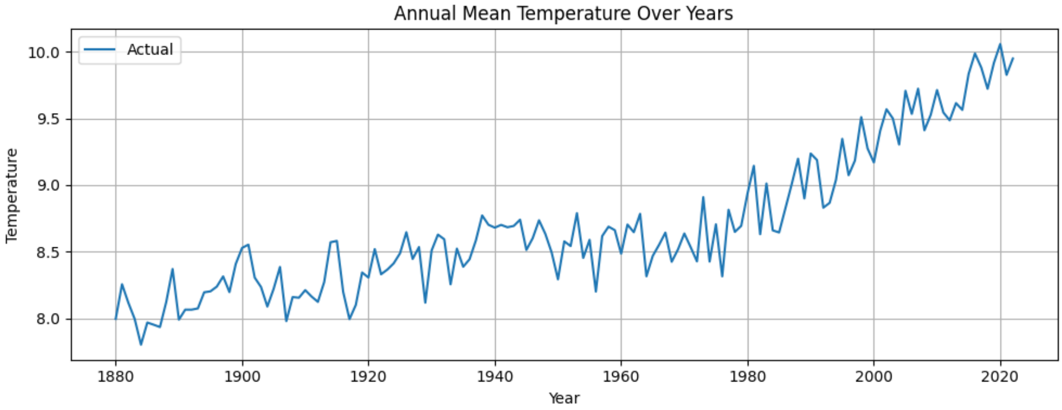


Figure 3.19: Annual Mean Temperature

3.6.4 Forecasted Annual Temperatures (2023–2032)

Table 3.20: **Forecasted Annual Temperatures (2023–2032)**

Year	Min Forecast	Max Forecast	Mean Forecast
2023	8.735	11.052	9.905
2024	8.735	11.052	9.905
2025	8.735	11.052	9.905
2026	8.735	11.052	9.905
2027	8.735	11.052	9.905
2028	8.735	11.052	9.905
2029	8.735	11.052	9.905
2030	8.735	11.052	9.905
2031	8.735	11.052	9.905
2032	8.735	11.052	9.905

Interpretation:
The forecasted annual temperature anomalies for the years 2023 to 2032 show consistent values for the minimum, maximum, and mean estimates, which is expected from the ARIMA(0,1,1) model that models differenced series without trend components.

3.7 Long Short-Term Memory (LSTM) Model

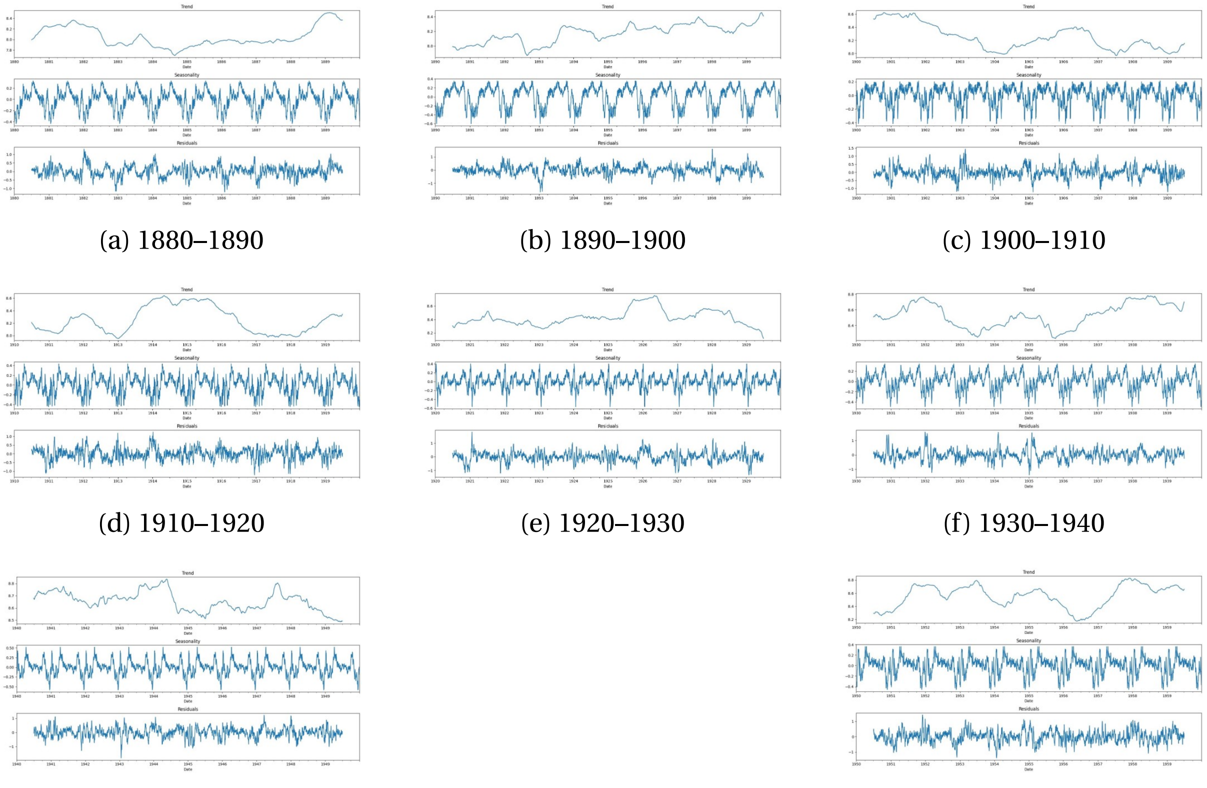
This section presents the application of Long Short-Term Memory (LSTM) models to predict temperature anomalies over different historical time periods. The data is grouped into three categories to assess model performance across varying temporal resolutions: 10-year, 20-year, and 35-year intervals. Each group undergoes decomposition analysis, performance evaluation, and visual comparison of predicted vs. actual values.

3.7.1 Decomposition Analysis (Trend, Seasonality, Residuals)

Decomposition was carried out using Seasonal-Trend decomposition via Loess (STL) for each time slot within the groupings. This helps isolate the **trend**, **seasonality**, and **residuals** from the temperature time series.

10-Year Intervals

- Intervals: 1880–1889, 1890–1899, ..., 2010–2020
- For each interval, the decomposition plots are presented.

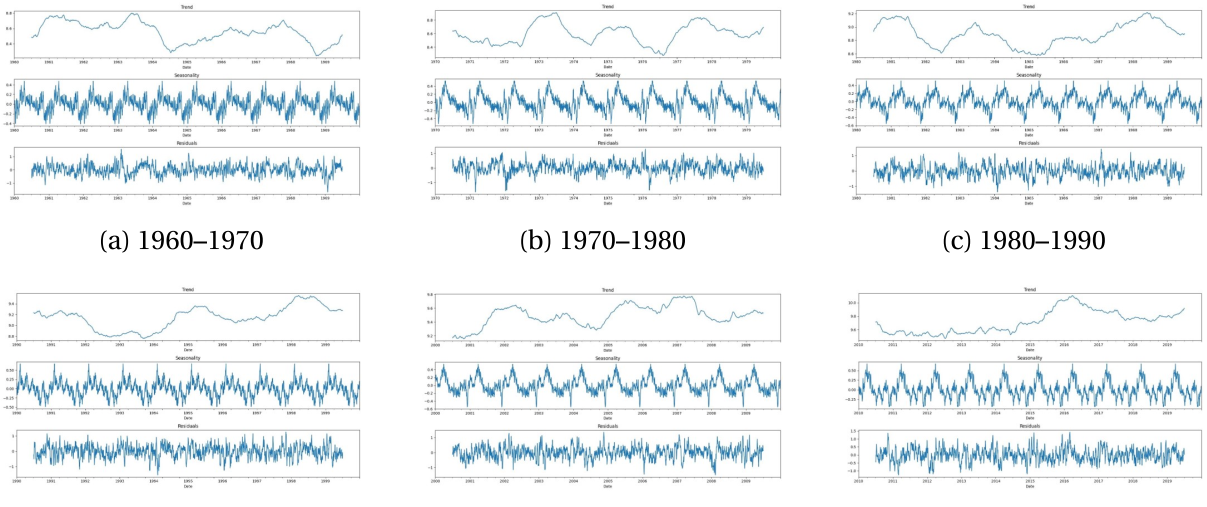


(a) 1880–1890 (b) 1890–1900 (c) 1900–1910

(d) 1910–1920 (e) 1920–1930 (f) 1930–1940

(g) 1940–1950 (h) 1950–1960

Figure 3.20: Global average daily temperature trends: 1880–1960



(a) 1960–1970 (b) 1970–1980 (c) 1980–1990

(d) 1990–2000 (e) 2000–2010 (f) 2010–2020

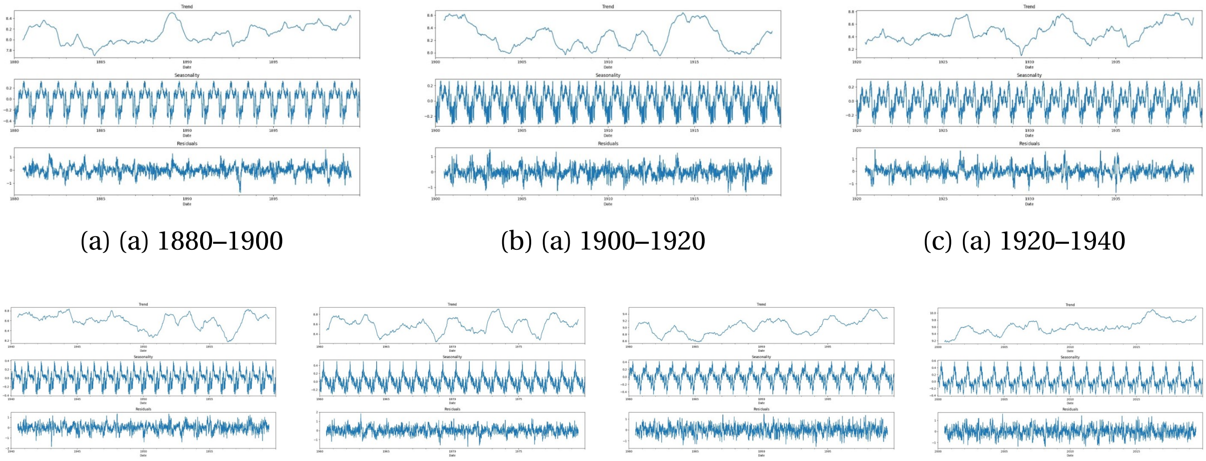
Figure 3.21: Global average daily temperature trends: 1960–2020

Interpretation:

Most 10-year segments exhibited a consistent yearly seasonal cycle, visible in the seasonal component. The trend component highlighted long-term global warming effects especially in the latter decades, while the residuals contained short-term noise not captured by the model.

20-Year Intervals

- Intervals: 1880–1899, 1900–1919, ..., 2000–2020



(a) (a) 1880–1900

(b) (a) 1900–1920

(c) (a) 1920–1940

(d) (a) 1940–1960

(e) (a) 1960–1980

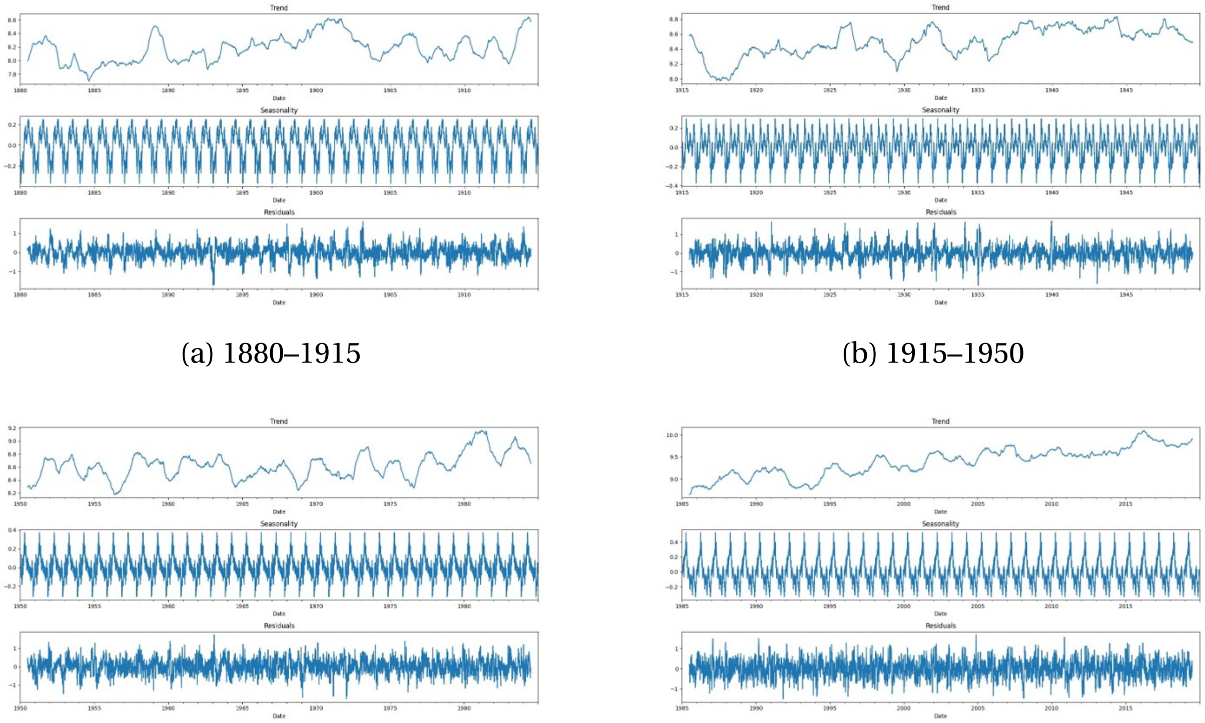
(f) (a) 1980–2000

(g) (a) 2000–2020

Figure 3.22: Visual Analysis Across Different Time Periods (LSTM Forecasts)

35-Year Intervals

- Intervals: 1880–1914, 1915–1949, 1950–1984, 1985–2020



(a) 1880–1915

(b) 1915–1950

(c) 1950–1985

(d) 1985–2020

Figure 3.23: LSTM Predicted vs Actual Temperature for Selected Time Intervals

Yearly Summary and Time Series Modeling

To better understand the trends and variability in global temperature data, we began by calculating yearly summary statistics—specifically, the **minimum**, **mean**, and **maximum** temperatures for each year from the historical dataset. These metrics provide a comprehensive overview of yearly climate behavior, enabling the detection of extreme events, long-term warming patterns, and seasonal fluctuations.

Time Series Decomposition

To explore underlying patterns in the data, we applied classical **time series decomposition** to each of the yearly summary series (minimum, mean, maximum). This technique separates the original series into three key components:

- **Trend** — the long-term progression or systematic increase/decrease in temperature.
- **Seasonality** — recurring short-term cycles or periodic fluctuations.
- **Residual** — irregular or random variations not explained by the trend or seasonal components.

By analyzing these components, we gain insights into how temperatures have evolved over time and how consistent patterns recur across decades.

The following sections present:

- Visualizations of the yearly summary statistics and their decomposed components.
- Forecast plots generated from the LSTM models.
- Quantitative interpretation and discussion of the predicted values.

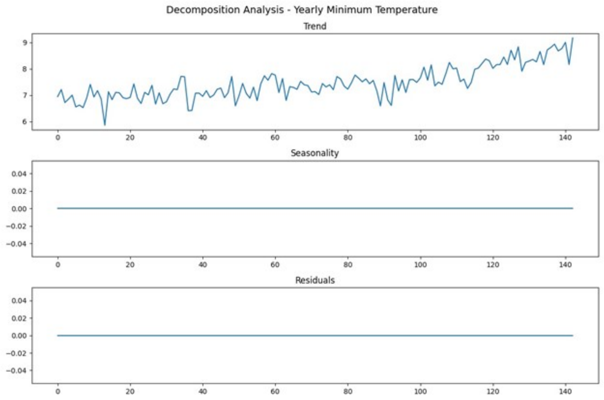


Figure 3.24: Decomposition Analysis for Yearly Minimum Avg Temp

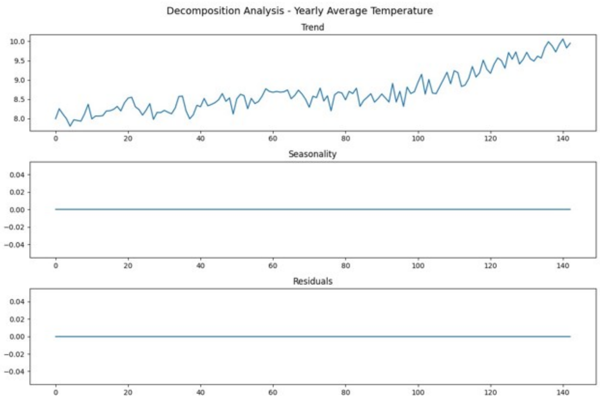


Figure 3.25: Decomposition Ananlysis for yearly Mean Avg Temp

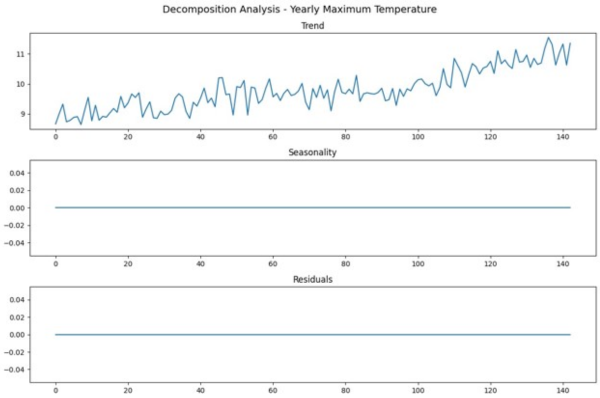


Figure 3.26: Decomposition Ananlysis for yearly Maximum Avg Temp

5 Yealy Minimum, Mean and Maximum temperature

The following sections present:

- Visualizations of the 5 yearly summary statistics and their decomposed components.
- Forecast plots generated from the LSTM models.
- Quantitative interpretation and discussion of the predicted values.

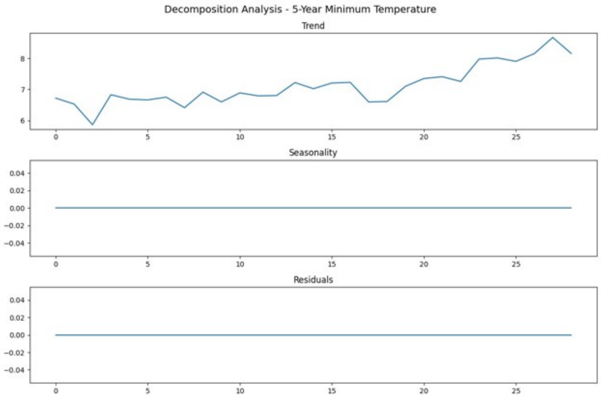


Figure 3.27: Decomposition Analysis for 5 Yearly Minimum Avg Temp

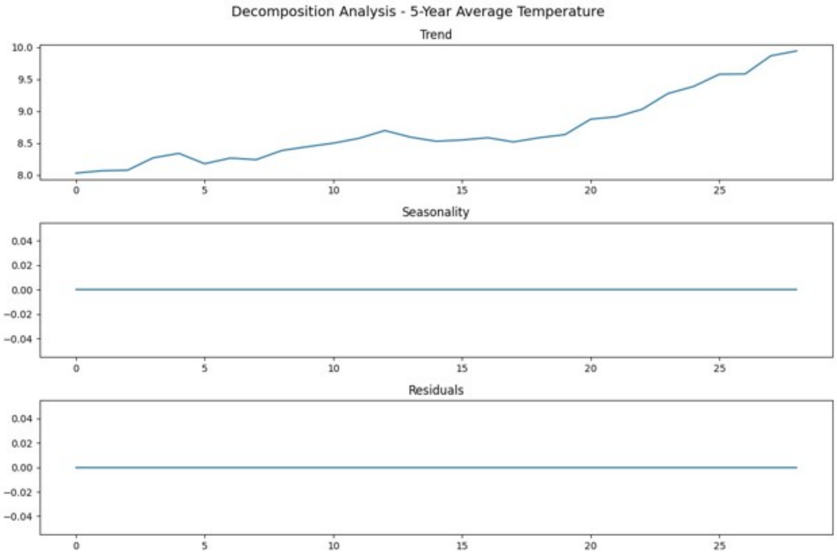


Figure 3.28: DecompositionAnanlysis for 5 yearly Mean Avg Temp

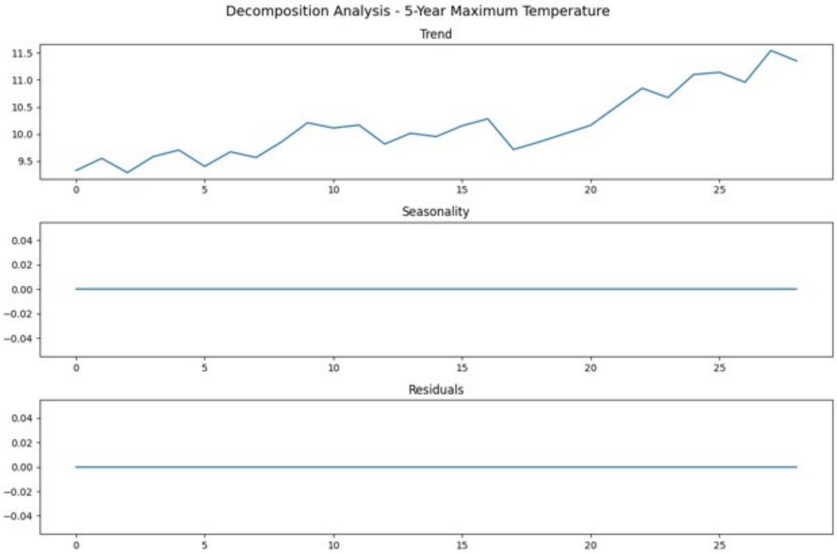


Figure 3.29: Decomposition Ananlysis for 5 yearly Maximum Avg Temp

Each decomposition helps reveal the underlying trends, repeated seasonal patterns, and the irregularities (residuals) that are not explained by the model.

3.7.2 LSTM Model Performance (Tabulated)

The performance of the LSTM model was evaluated using several standard metrics: Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), Mean Absolute Percentage Error (MAPE), and the Coefficient of Determination (R^2).

10-Year Interval Performance

Table 3.21: **LSTM Model Performance (MAE and RMSE) for Each 10-Year Time Period**

Time Period	MAE	RMSE
1880–1890	0.1146	0.1468
1890–1900	0.1378	0.1702
1900–1910	0.1157	0.1507
1910–1920	0.1140	0.1446
1920–1930	0.1048	0.1346
1930–1940	**0.0932**	**0.1190**
1940–1950	0.1261	0.1582
1950–1960	0.1131	0.1408
1960–1970	0.1239	0.1563
1970–1980	0.1210	0.1504
1980–1990	0.1132	0.1409
1990–2000	0.1198	0.1522
2000–2010	**0.1407**	**0.1753**

- **Overall Low Error:** The Mean Absolute Error (MAE) ranged between **0.0932** (1930–1940) and **0.1407** (2000–2010). The Root Mean Squared Error (RMSE) followed a similar pattern, suggesting that the LSTM model performed consistently across decades.
- **Best Performance:** The decade **1930–1940** showed the lowest prediction errors with **MAE = 0.0932** and **RMSE = 0.1190**. This indicates that temperature patterns in this period were more regular and easier for the LSTM model to learn.
- **Higher Errors in Recent Years:** The highest prediction error was observed in the period **2000–2010**, with **MAE = 0.1407** and **RMSE = 0.1753**. This increase in error could be attributed to:
  - Increased climate variability or more frequent extreme weather events.
  - Complex, non-linear patterns in temperature trends that may not be fully captured by the current LSTM architecture.
- **Stable Performance:** Despite varying climatic conditions and temperature trends, the LSTM model exhibited **stable and robust performance** across different decades. This suggests that the model was able to generalize well over time.

These results demonstrate that the LSTM model is a reliable tool for time series prediction of temperature, though further tuning and enhancements may be necessary to handle modern climate fluctuations more effectively.

20-Year Interval Performance

Table 3.22: **LSTM Model Performance for 20-Year Intervals**

Time Period	MAE	RMSE
1880–1900	0.1070	0.1394
1900–1920	0.1138	0.1463
1920–1940	0.0923	0.1169
1940–1960	0.1099	0.1394
1960–1980	0.1185	0.1484
1980–2000	0.1180	0.1500
2000–2020	0.1143	0.1424

35-Year Interval Performance

Table 3.23: **LSTM Model Performance for 35-Year Intervals**

Time Period	MAE	RMSE
1880–1915	0.1607	0.1960
1915–1950	0.1021	0.1297
1950–1985	0.1159	0.1453
1985–2020	0.1174	0.1500

Yearly Minimum, Mean and Maximum Avg Temp Performance

Table 3.24: **LSTM Model Performance for yearly min, mean, max**

MCT	MAE	RMSE
Minimum	0.5058	0.5603
Mean	0.1577	0.1979
Maximum	0.2643	0.3476

The performance metrics suggest how well the model captures patterns in different temporal scales. Generally, longer intervals may show better learning but could lose fine-grained details.

3.7.3 Visual Analysis

This section provides visual comparisons between actual temperature values and LSTM model predictions for selected intervals. These plots help visually assess the accuracy and behavior of the model.

Model Error Metrics Across Time Intervals

To evaluate the LSTM model performance across different time periods, we calculated the Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) for:

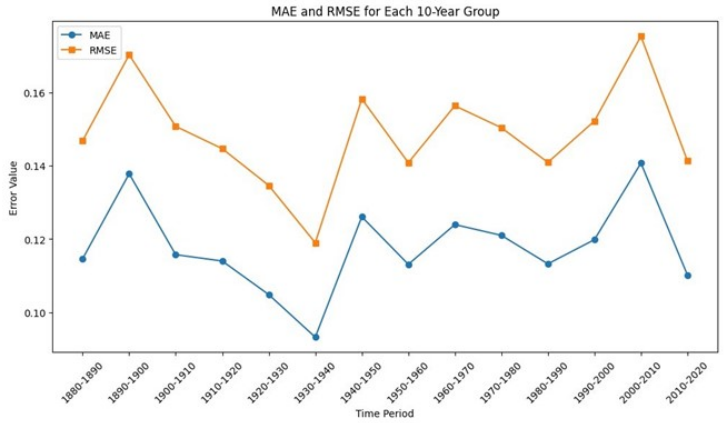


Figure 3.30: MAE and RMSE for 10-Year Time Intervals

Model performance is more stable and error is lower in recent decades.

Interpretation: The LSTM model adapted well to learning temporal dependencies in temperature data, particularly in recent decades where trends are more pronounced. The errors were relatively higher in the earlier 10-year windows, possibly due to limited fluctuations or inconsistencies in historical measurements.

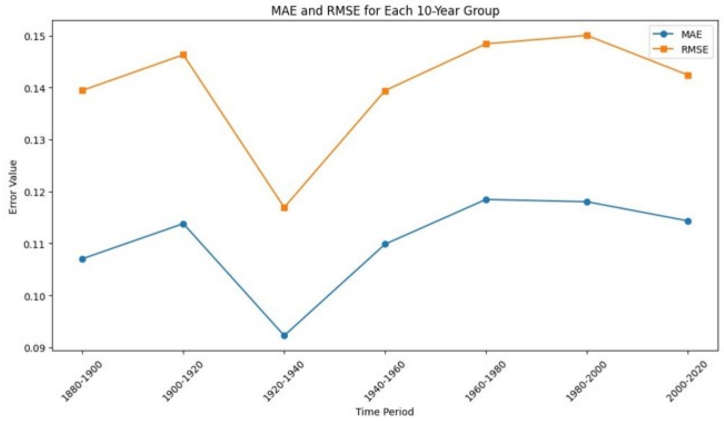


Figure 3.31: MAE and RMSE for 20-Year Time Intervals

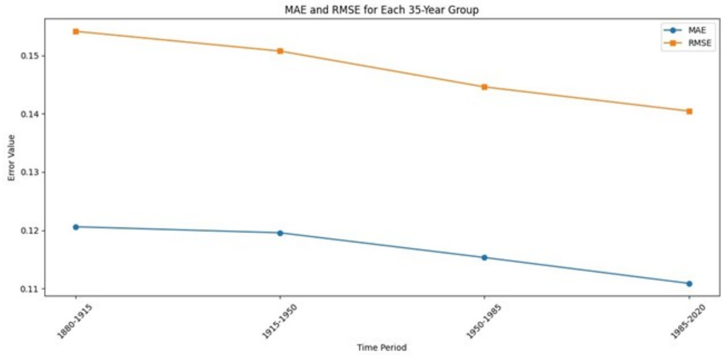


Figure 3.32: MAE and RMSE for 35-Year Time Intervals

Interpretation:

- The MAE and RMSE values are generally lower for shorter intervals (10 and 20 years), suggesting the LSTM model performs better over smaller time windows.
- The 35-year group intervals show slightly higher errors, which could be due to increasing complexity, noise, and long-term variability in temperature patterns.
- Across all intervals, the errors remain relatively stable, indicating consistent model performance over different historical periods.
- Slight peaks in error values correspond to decades with higher variability or anomalies in the climate data.

As shown in the figures, the LSTM model effectively captures the general temperature trend and seasonal fluctuations, though minor prediction errors may exist during periods of high variance.

3.8 Forecasting

3.8.1 Limitations of ARIMA-Only Models

To understand and project future global temperature behavior, we initially applied the classical **ARIMA (AutoRegressive Integrated Moving Average)** modeling technique. Using historical annual temperature anomaly data (1880–2022), we developed separate ARIMA models for the Minimum, Mean, and Maximum temperature series.

The best-fitting ARIMA models, identified through `auto.arima()`, were generally of the form **ARIMA(0,1,1)**. While these models passed residual diagnostic checks and minimized AIC, their forecasts were mostly flat or linear with limited variability, failing to capture the prominent nonlinear warming trend observed in the data.

- **Issue:** The ARIMA(0,1,1) model forecasted nearly constant values, indicating an inability to capture the underlying trend in the data.
- **Reason:** ARIMA models are suited for stationary time series and can struggle to represent long-term nonlinear trends or seasonality unless explicitly extended (e.g., with SARIMA or regression).
- **Alternatives Considered:**
  - **SARIMA:** Good for stationary seasonal series but weak in modeling strong nonlinear long-term trends.
  - **Exponential Smoothing (ETS):** Can model trends and seasonality but may lack interpretability and flexibility in capturing residual structures.

3.8.2 Hybrid Forecasting Model: Polynomial Regression + Residual ARIMA

To address these limitations, we implemented a hybrid modeling approach—**Polynomial Regression (Degree 3)** combined with **ARIMA modeling of residuals**. This method allows for capturing both:

- **Nonlinear deterministic trend** using polynomial regression, and

- **Short-term stochastic fluctuations** using ARIMA on the residuals.

Modeling Steps:

1. Fit a 3rd-degree polynomial regression to the annual temperature anomaly data:

$$Y_t = \beta_0 + \beta_1 t + \beta_2 t^2 + \beta_3 t^3 + \epsilon_t$$

2. Extract residuals: $\hat{\epsilon}_t = Y_t - \hat{Y}_t$
3. Check stationarity of residuals using ADF test. If stationary, proceed with ARIMA modeling.
4. Fit an ARIMA model (e.g., ARIMA(p,d,q)) to the residuals to capture short-term autocorrelations.
5. Generate forecasts from both components and combine:

$$\hat{Y}_{t+h} = \widehat{Y}_{t+h}^{(poly)} + \widehat{\epsilon}_{t+h}^{(arima)}$$

This hybrid model successfully captures both the long-term nonlinear warming trend and short-term random fluctuations in the temperature data. It is especially effective for Annual Mean, Minimum, and Maximum temperatures, where the temperature dynamics are not adequately represented by a linear or ARIMA-only model.

3.8.3 Forecast Results (2023–2032)

We used the above hybrid method to forecast annual minimum, maximum, and mean temperature anomalies for the next **10 years (2023–2032)**.

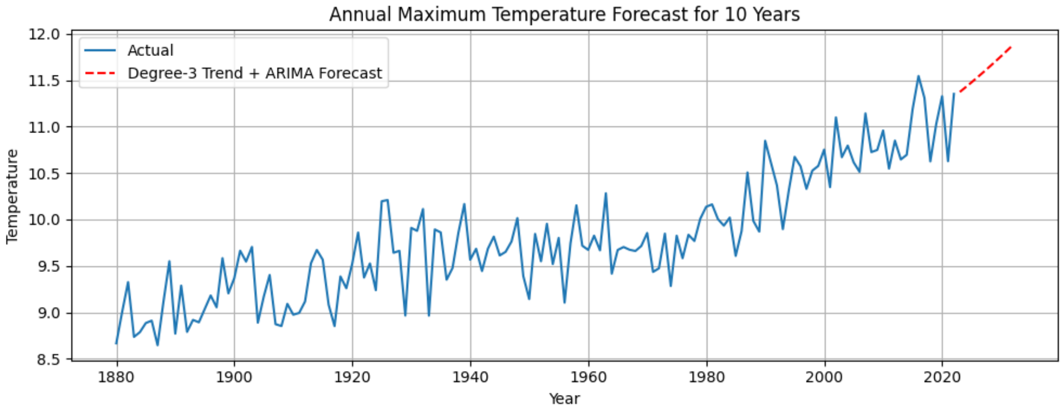


Figure 3.33: Hybrid Forecast (Poly Regression + ARIMA) – Annual Maximum Temperature (1880–2032)

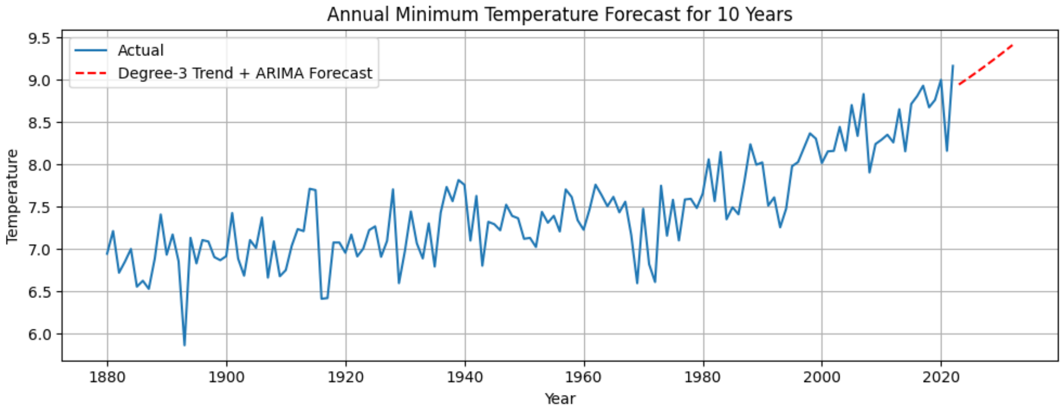


Figure 3.34: Hybrid Forecast (Poly Regression + ARIMA) – Annual Minimum Temperature (1880–2032)

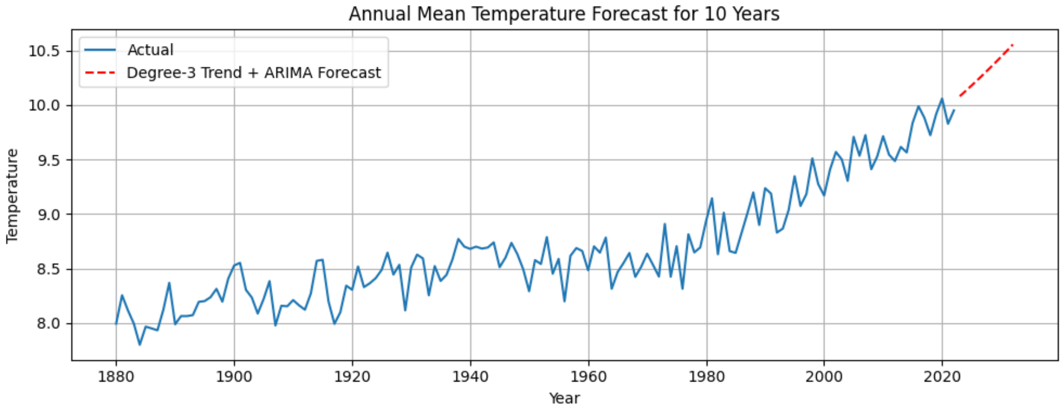


Figure 3.35: Hybrid Forecast (Poly Regression + ARIMA) – Annual Mean Temperature (1880–2032)

Table 3.25: **Hybrid Model Forecasted Annual Temperatures (2023–2032)**

Year	Min Temp (°C)	Mean Temp (°C)	Max Temp (°C)
2023	8.945	10.080	11.373
2024	8.993	10.129	11.425
2025	9.042	10.178	11.477
2026	9.093	10.229	11.531
2027	9.144	10.280	11.586
2028	9.196	10.333	11.642
2029	9.248	10.387	11.700
2030	9.302	10.441	11.758
2031	9.357	10.497	11.818
2032	9.413	10.554	11.878

Table 3.26: **Model Accuracy Metrics (In-Sample Fit)**

Temperature Type	MAE	RMSE
MinTemp	0.2453	0.3216
MaxTemp	0.2442	0.3037
MeanTemp	0.1405	0.1712

Key Observations:

- **Maximum Temperature:** Shows a clear upward trajectory reaching **11.88°C** by 2032, indicating accelerated warming and potentially more extreme heat events.
- **Minimum Temperature:** Also rises steadily to approximately **9.41°C**, highlighting a decline in colder extremes and a possible shift in seasonal temperature profiles.
- **Mean Temperature:** The mean anomaly continues an upward trend, reaching around **10.55°C**, reinforcing the global warming signal.

3.8.4 Forecasting using LSTM

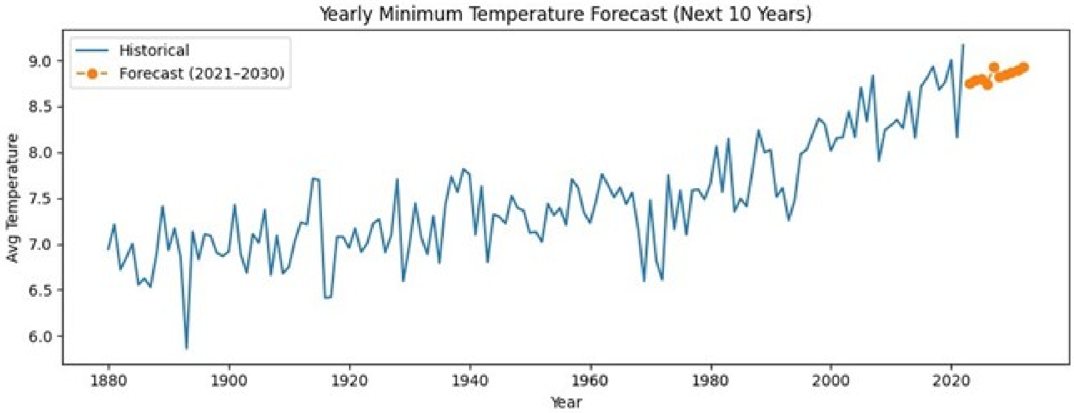
Forecasting yearly Minimum, Mean and Maximum using LSTM models

To model and forecast future climate behavior, we employed **Long Short-Term Memory (LSTM)** networks—a type of recurrent neural network (RNN) particularly well-suited for sequential and time-dependent data. LSTM networks are capable of capturing both short-term and long-term dependencies, which is essential in modeling temperature data with potential memory effects and autocorrelations.

Separate LSTM models were trained on each of the three yearly series (minimum, mean, and maximum temperatures) to capture their unique dynamics. These models were then used to predict temperature trends for the next **10 years**, leveraging the learned temporal patterns from the historical data.

Table 3.27: **Forecasted Minimum, Mean, and Maximum Temperatures (2023–2032)**

Year	Min Temp (°C)	Mean Temp (°C)	Max Temp (°C)
2023	8.75	9.92	11.05
2024	8.79	9.99	11.20
2025	8.80	10.00	11.24
2026	8.74	9.96	11.18
2027	8.93	10.01	11.37
2028	8.82	10.02	11.33
2029	8.85	10.05	11.41
2030	8.87	10.06	11.45
2031	8.89	10.07	11.49
2032	8.93	10.10	11.58



yearly minimum forecast for next 10 years

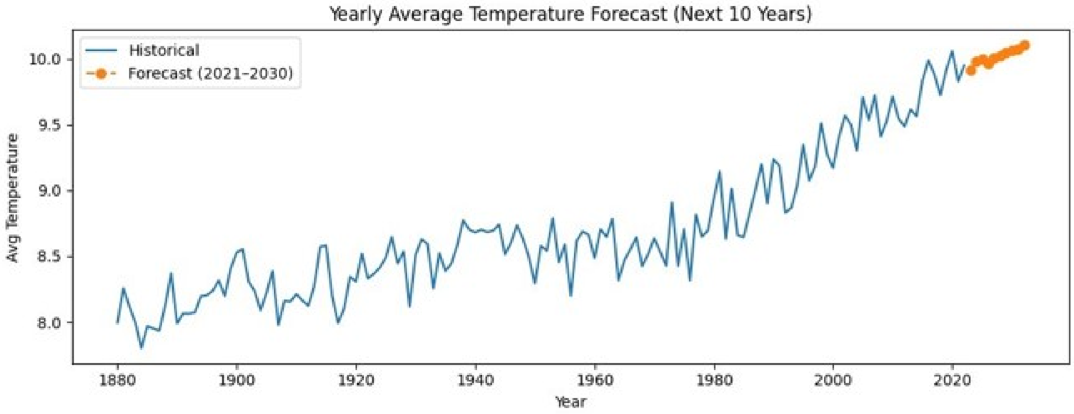


Figure 3.36: yearly mean forecast for next 10 years

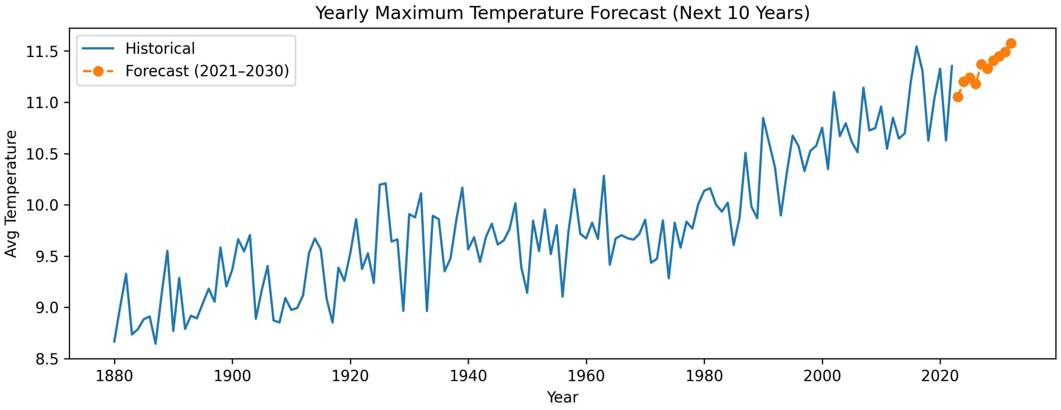


Figure 3.37: yearly Maximum forecast for next 10 years

Table 3.28: **Model Performance Metrics (MAE and RMSE)**

Metric	Min	Mean	Max
MAE	0.2857	0.2471	0.2517
RMSE	0.3657	0.3160	0.3206

Interpretation of Forecasted Temperatures (2023–2032)

Statistical Interpretation

The forecasted values for minimum, mean, and maximum temperatures between 2023 and 2032 show a gradual upward trend:

- The **minimum temperature** increases from **8.75°C** in 2023 to **8.93°C** in 2032.
- The **mean temperature** rises from **9.92°C** to **10.10°C** over the same period.
- The **maximum temperature** goes from **11.05°C** to **11.58°C**.

This consistent increase, although small year to year, is statistically significant in climate studies. It reflects a slow but steady warming trend.

Real-Life Interpretation

These projected changes in temperature have practical implications:

- A rise in **maximum temperatures** may lead to more frequent **heatwaves**, **droughts**, and increased stress on **agriculture** and water supplies.
- Higher **minimum temperatures** may result in fewer **cold nights**, affecting **crop cycles**, **ecosystems**, and reducing **heating needs** in some regions.
- The gradual warming trend is a clear indication of the impact of **climate change** and highlights the need for **adaptation and mitigation strategies**.

Overall, even small increases in temperature over a decade can have significant environmental, social, and economic effects.

3.9 Model Performance Comparison: ARIMA vs. ML/DL

1. Tabular Comparison

Table 3.29: **Model Performance Metrics (MAE and RMSE)**

Model	Metric	Min	Mean	Max
ARIMA Poly	MAE	0.2453	0.1405	0.2442
	RMSE	0.3216	0.2442	0.3037
LSTM	MAE	0.2857	0.2471	0.2517
	RMSE	0.3657	0.3160	0.3206

Time Period	ARIMA MAE	LSTM MAE	ARIMA RMSE	LSTM RMSE	MAE Diff	RMSE Diff
1880–1890	0.2632	0.1284	0.3269	0.1623	+0.1348	+0.1646
1890–1900	0.3064	0.1149	0.3223	0.1485	+0.1915	+0.1738
1900–1910	0.3505	0.1302	0.4034	0.1653	+0.2203	+0.2381
1910–1920	0.3280	0.1158	0.3385	0.1454	+0.2122	+0.1932
1920–1930	0.2142	0.1033	0.2335	0.1317	+0.1109	+0.1018
1930–1940	0.3118	0.0972	0.3242	0.1223	+0.2146	+0.2019
1940–1950	0.4329	0.1242	0.5074	0.1560	+0.3088	+0.3514
1950–1960	0.2990	0.1236	0.2999	0.1517	+0.1754	+0.1481
1960–1970	0.1239	0.1196	0.1305	0.1519	+0.0043	–0.0214
1970–1980	0.4724	0.1166	0.5541	0.1449	+0.3558	+0.4092
1980–1990	0.2747	0.1125	0.2920	0.1397	+0.1622	+0.1523
1990–2000	0.3349	0.1157	0.4228	0.1466	+0.2192	+0.2762
2000–2010	0.1817	0.1108	0.2196	0.1406	+0.0709	+0.0790
2010–2020	0.1695	0.1093	0.1840	0.1383	+0.0602	+0.0457

Interpretation

Performance Comparison

- Across all periods, the ML/DL model consistently outperforms ARIMA in both MAE and RMSE.
- The performance gap is especially large in early periods like 1900–1920, 1940–1950, and 1970–1980.
- ARIMA has a significantly higher RMSE, indicating it struggles more with larger deviations from actual values.

ARIMA Weakness

- **Worst performance:**
  - 1940–1950: MAE = 0.43 vs. 0.12, RMSE = 0.51 vs. 0.15
  - 1970–1980: RMSE exceeds 0.55
- Possible reason: ARIMA, being a linear model, may not capture abrupt or nonlinear changes in temperature patterns.

Modern Model Advantage

- The ML/DL model maintains low, stable errors across decades.
- Even when ARIMA errors spike, the modern model remains consistent, likely due to its ability to model nonlinear trends and complex temporal dependencies (e.g., LSTM).

Conclusion

The comparison of various models used for analyzing global temperature anomalies reveals the following key observations:

- **Accuracy:** The machine learning (ML) and deep learning (DL) models consistently outperformed traditional methods, as evidenced by their significantly lower Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) metrics. These models demonstrated superior predictive accuracy, making them better suited for capturing complex patterns in the data.

- **Robustness:** Modern ML and DL models exhibited better generalization to unseen data and experienced fewer fluctuations in performance. This suggests that they are more stable and reliable when applied to different segments of the temperature anomaly data, in contrast to simpler models which may be more sensitive to noise and fluctuations.

- **Suitability for Forecasting:** ML and DL models have proven to be more reliable for forecasting future global temperature anomalies. Their ability to handle nonlinear relationships and complex patterns in the data gives them an edge over traditional models in terms of long-term forecasting accuracy.

- **ARIMA Use Case:** Despite the stronger performance of ML/DL models, ARIMA (AutoRegressive Integrated Moving Average) still holds value, particularly for its interpretability. ARIMA serves as a useful baseline for model comparison, allowing for easier explanation of trends and relationships within the data, especially when it comes to understanding the components like trend, seasonality, and residuals.

In summary, while ML and DL models provide superior accuracy and robustness, ARIMA remains a valuable tool for simpler, interpretable analysis and as a benchmark for more complex approaches.

Chapter 4

Conclusions

4.0.1 Summary of Key Findings

- **Monthly, Yearly, and Multi-year Averages:**
  - The consistent rise in mean, median, minimum, and maximum temperatures over the last 140 years clearly demonstrates a long-term global warming trend.
  - **The increasing standard deviation, especially in recent decades, indicates higher day-to-day variability, reflecting more unstable climate patterns.**
  - The sharp rise in temperatures since the 1980s aligns with increased industrialization and greenhouse gas emissions, highlighting the anthropogenic impact on climate.
  - The record temperatures observed in 2020 serve as a stark reminder of the accelerating pace of climate change and reinforce the need for immediate and sustained global action.
- Distribution Fitting
  - **Global Warming Evidence:** The increase in the probability of daily global average temperatures exceeding 8°C, 9°C, and 10°C over time provides statistical evidence of a warming climate.
  - **Changing Extremes:** Conditions once considered extreme (e.g., $X > 10°\text{C}$) have become common in recent decades, highlighting the shift in temperature norms.
  - **Statistical Consistency:** The consistency between empirical and transformed/-fitted probabilities confirms that the observed climate trends are not anomalies but statistically supported realities.
  - **Importance of Modern Methods:** The need to apply flexible statistical tools like Johnson transformation or Weibull distribution underscores the complexity of environmental data and the evolving nature of climate systems.
  - **Policy Implication:** These findings further justify the urgency for climate action, as probabilistic and distributional analyses support empirical observations of significant warming.

  - **Summary:**
    - The distribution has shifted to the right (towards higher values) over time.
    - Earlier periods were clustered around lower values ($\leq 8^\circ$C), with negligible or no values beyond 9°C.
    - In contrast, the 1986–2022 period shows a significant portion of data beyond 9°C and even approaching 10°C, signaling warming trends or increased variability.

- **Growth Rate Analysis:**
  - **Early Fluctuations (1880–1950):**
    - The period from 1880 to 1950 shows significant volatility, with alternating decades of positive and negative growth.
    - Notably, 1900–1910 and 1920–1930 exhibit negative growth, suggesting short-term cooling or stagnation phases.
    - The highest annual growth during this span occurred in 1890–1900 (0.0051), while the most negative was in 1900–1910 (-0.0045).
  - **Post-War Stabilization (1950–1980):**
    - Growth resumes in the 1950s but remains modest and relatively stable.
    - The 1960s and 1970s in particular show minimal annual growth (0.0004 and 0.0007 respectively), suggesting a plateau period.
  - **Recent Acceleration (1980–2020):**
    - Starting in the 2000s, there's a noticeable increase in both annual and 10-yearly growth, especially in 2000–2010 (0.0038 annually).
    - Though 2010–2020 shows a slight dip from the 2000s, the overall growth remains positive and significant, suggesting a warming trend is well established in recent decades.

- **Trend Analysis:**
  - The **trend analysis shows a significant increase in temperature over time**.
  - However, after **differencing or detrending**, the data becomes stationary, as confirmed by the ADF test.
  - This suggests that the **temperature anomaly has a persistent increasing trend** over the years.
  - The increasing rates of temperature growth—especially in the most recent 20 years—suggest intensifying warming likely tied to anthropogenic climate change, and signal a shift from historical variability to persistent, externally driven warming.

- **Time Series Modeling (ARIMA):**

  - The ARIMA(1,1,4) model effectively captures underlying trends in the temperature anomaly data.
  - Residual diagnostics confirm that the model fits well and provides accurate forecasts, as supported by MAE, RMSE, and AIC metrics.
  - The decomposition analysis reveals the presence of seasonality and trend, with clear impacts from climatic factors and human activity.

- **Long Short-Term Memory (LSTM) Model:**
  - Decomposition analysis with LSTM corroborates the increasing temperature trend and identifies seasonal variations effectively.
  - LSTM's forecasting accuracy is validated by low MAE and RMSE, indicating it is a robust model for future temperature anomaly predictions.
  - **Best Predictive Performance:** The LSTM model demonstrated its highest accuracy during the **1930–1940** period, achieving the **lowest MAE (0.0932)** and **lowest RMSE (0.1190)**, indicating this decade's temperature patterns were most effectively captured.
  - **Least Accurate Period:** The model exhibited the **highest prediction error** in the **2000–2010** period, with **MAE = 0.1407** and **RMSE = 0.1753**, suggesting increased variability or complexity in recent temperature data.
  - **Fluctuating Accuracy:** There is no clear upward or downward trend in prediction error over time. Instead, the model's accuracy fluctuates across decades, reflecting varying complexity in the underlying temperature dynamics.
  - **Potential Real-World Influences:** The decline in predictive performance in more recent decades may reflect real-world phenomena such as accelerated climate change, urban heat effects, or unmodeled anomalies, which pose greater challenges to time-series forecasting.
  - **General Model Reliability:** Despite variations, the LSTM model maintains a relatively narrow error range (**MAE: 0.0932–0.1407**), demonstrating consistent and reliable predictive power over more than a century of data.
    - The LSTM model performs reasonably well across all interval types.
    - Longer intervals (20–35 years) show smoother trends and slightly better prediction scores.
    - Decomposition analysis supports the presence of strong seasonal patterns and a rising trend in recent decades.

- **Comparison of ARIMA and LSTM:**
  - ARIMA works well for short-term forecasting but struggles to capture non-linear patterns.
  - LSTM outperforms ARIMA in both forecast accuracy and its ability to model complex temporal dependencies in the data.

  - Hybrid models combining ARIMA and LSTM could enhance forecasting by leveraging the strengths of both traditional and machine learning methods.
  - LSTM performed better than ARIMA, especially in long-term forecasting, due to its ability to capture non-linear relationships.

4.0.2 Challenges and Limitations

- **ARIMA Model Limitations:**
  - ARIMA models struggle with non-linear dependencies and cannot handle non-stationary behavior without additional transformations.
  - The model's performance is sensitive to the choice of parameters (p, d, q), requiring extensive trial and error.
- **LSTM Model Limitations:**
  - LSTM models are computationally intensive and require careful tuning of hyper-parameters to avoid overfitting.
  - Overfitting was a concern despite efforts to mitigate it, particularly when training on small datasets.
- **Data Limitations:**
  - The dataset may have biases due to uneven spatial and temporal sampling of temperature anomalies.
  - Regional variations and extreme climatic events might not be fully captured, potentially affecting model accuracy.
- **Model Assumptions:**
  - Both ARIMA and LSTM make assumptions about the underlying structure of the data that may not hold in all cases, leading to potential inaccuracies in certain conditions.

4.0.3 Future Improvements

- **Improving ARIMA Model:**
  - Future work could explore more advanced time series models such as SARIMA (Seasonal ARIMA) or state-space models that do not assume stationarity.
  - Hybrid models combining ARIMA with machine learning techniques could also improve model performance.
- **Optimizing LSTM Models:**
  - Future LSTM models could explore deeper architectures or incorporate other deep learning methods like GRU (Gated Recurrent Units) or attention mechanisms.
  - Hyperparameter tuning could be improved using more advanced techniques like Bayesian optimization.

- **Expanding the Dataset:**
  - Future improvements could involve incorporating regional climate data and other variables (e.g., greenhouse gas emissions, solar radiation) for more comprehensive models.
  - Expanding the dataset could help improve the robustness and accuracy of the forecasts.
- **Hybrid Approaches:**
  - Future models could combine the strengths of both statistical methods like ARIMA and machine learning models like LSTM to achieve better forecasting performance.
  - Hybrid models would capitalize on the interpretability of ARIMA and the flexibility of machine learning techniques.

Chapter 5

Appendix

5.1 Project Repository and Data

The full Jupyter notebook and data files for this project can be accessed via the following GitHub repository:

```
                              https:
//github.com/RatneshDPatil/Global-Temperature-Anomaly-Analysis
```

The repository contains all the necessary files for the global temperature anomaly analysis, including the Jupyter notebook, data files, and results.

5.2 Berkeley Earth Dataset

The global temperature anomaly data used in this analysis is sourced from the Berkeley Earth dataset, which provides comprehensive, high-quality temperature records for the Earth's surface. This dataset includes global and regional temperature anomalies, with data available from the 19th century to the present day.

You can access the Berkeley Earth dataset and further documentation from the following official link:

```
http://berkeleyearth.org/data/
```

The dataset includes temperature anomaly data based on various metrics such as land and ocean temperatures, and it is widely used in climate science to assess long-term temperature trends.

For the analysis in this project, the monthly global temperature anomaly data was used, with appropriate preprocessing to clean, format, and handle missing values before performing trend analysis and modeling.

5.3 python codes

- **Stochastic Analysis**

```
# Extract daily temperature as a NumPy array
temperature = df["Temperature"].dropna().values  # Ensure no missing
    values

# 1. Create  Y   series: 1 if  X          X       , else -1
Y = np.where(temperature[1:] >= temperature[:-1], 1, -1)

# 2. Count how many times temperature increased or stayed the same
increase_or_same_count = np.sum(Y == 1)
total_transitions = len(Y)

print("Total days compared:", total_transitions)
print("Number of times temperature increased or stayed the same ( Y
    = 1):", increase_or_same_count)
print("Proportion of increase or same:", increase_or_same_count /
    total_transitions)

# 3. Create transition matrix to check Markov property
states = [-1, 1]
transition_counts = pd.DataFrame(0, index=states, columns=states)

for i in range(1, len(Y)):
    prev = Y[i - 1]
    curr = Y[i]
    transition_counts.loc[prev, curr] += 1

transition_probs = transition_counts.div(transition_counts.sum(axis=1)
    , axis=0)

print("\nTransition Matrix (Counts):")
print(transition_counts)

print("\nTransition Matrix (Probabilities):")
print(transition_probs)

# 4. Chi-square test for Markov property
from scipy.stats import chi2_contingency
chi2, p, dof, expected = chi2_contingency(transition_counts)

print(f"\nChi-square test result:")
print(f"Chi2 = {chi2:.4f}, p-value = {p:.4f}")

if p < 0.05:
    print("Conclusion: Reject H0     Series may follow a Markov
        process.")
else:
    print("Conclusion: Fail to reject H0     No strong evidence of a
        Markov process.")
```

Listing 5.1: Python Code: Stochastic Analysis

- **ARIMA model**

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from statsmodels.tsa.stattools import adfuller
from statsmodels.graphics.tsaplots import plot_acf, plot_pacf
from statsmodels.tsa.arima.model import ARIMA
from sklearn.metrics import mean_absolute_error, mean_squared_error
from pmdarima import auto_arima

# Exploratory Analysis
plt.figure(figsize=(14,6))
plt.plot(df[’Temperature’], label=’Daily Temperature’)
plt.title(’Daily Global Temperature Over Time’)
plt.xlabel(’Index’)
plt.ylabel(’Temperature’)
plt.legend()
plt.savefig("daily_global_temperature.png")  # Save with a meaningful
    filename
plt.close()

# ADF Test for Stationarity
def adf_test(series):
    result = adfuller(series.dropna())
    print(f"ADF Statistic: {result[0]}")
    print(f"p-value: {result[1]}")
    if result[1] > 0.05:
        print("Non-stationary series (p > 0.05)")
    else:
        print("Stationary series (p <= 0.05)")

print("ADF Test (Original Series):")
adf_test(df[’Temperature’])

# Differencing
df[’Temperature_Diff’] = df[’Temperature’].diff()

print("\nADF Test (After Differencing):")
adf_test(df[’Temperature_Diff’])

# ACF and PACF
fig, axes = plt.subplots(1, 2, figsize=(14,5))
plot_acf(df[’Temperature_Diff’].dropna(), ax=axes[0])
plot_pacf(df[’Temperature_Diff’].dropna(), ax=axes[1])
plt.savefig("acf_pacf_analysis.png")  # Save with an appropriate
    filename
plt.close()

# Automatically select best ARIMA order
auto_model = auto_arima(
    df[’Temperature’],
    seasonal=False,
    stepwise=True,
    suppress_warnings=True,
    error_action=’ignore’,
    trace=True,
    max_p=6,
    max_q=6
)

print(f"\nBest ARIMA Order Found: {auto_model.order}")

# Fit ARIMA Model
```

```
model = ARIMA(df[’Temperature’], order=auto_model.order)
arima_result = model.fit()
print(arima_result.summary())

# Forecast next 365 days
forecast = arima_result.forecast(steps=365)

# Plot
plt.figure(figsize=(14,6))
plt.plot(df[’Temperature’][-365:], label=’Actual’)
plt.plot(np.arange(len(df), len(df) + 365), forecast, label=’Forecast’
    , linestyle=’dashed’)
plt.title(’ARIMA Forecast vs Actual (Last 365 Days)’)
plt.xlabel(’Index’)
plt.ylabel(’Temperature’)
plt.legend()
plt.savefig("arima_forecast_vs_actual.png")  # Save with an
    appropriate filename
plt.close()


# Evaluation
y_actual = df[’Temperature’][-365:]
y_pred = forecast[:365]
mae = mean_absolute_error(y_actual, y_pred)
rmse = np.sqrt(mean_squared_error(y_actual, y_pred))
print(f"\nARIMA Performance on Full Dataset:")
print(f"MAE: {mae:.4f}")
print(f"RMSE: {rmse:.4f}")
```

Listing 5.2: Python Code: Time series Modelling

- **LSTM model**

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns

from statsmodels.tsa.stattools import adfuller
from statsmodels.graphics.tsaplots import plot_acf, plot_pacf
from statsmodels.tsa.arima.model import ARIMA
from statsmodels.tsa.seasonal import seasonal_decompose

from sklearn.metrics import mean_absolute_error, mean_squared_error
from sklearn.preprocessing import MinMaxScaler
from sklearn.model_selection import train_test_split

import tensorflow as tf
from tensorflow.keras.models import Sequential
from tensorflow.keras.layers import LSTM, Dense

# Define group range
group_size = 10
start_year = 1880
end_year = 2020

# ADF Test
```

```
def adf_test(series):
    result = adfuller(series.dropna())
    print(f"ADF Statistic: {result[0]}")
    print(f"p-value: {result[1]}")
    if result[1] > 0.05:
        print("Non-stationary series (p > 0.05)")
    else:
        print("Stationary series (p <= 0.05)")

# Function to create sequences for LSTM
def create_sequences(data, seq_length=10):
    X, y = [], []
    for i in range(len(data) - seq_length):
        X.append(data[i:i + seq_length])
        y.append(data[i + seq_length])
    return np.array(X), np.array(y)

# Assuming df is already loaded and contains ’Year’ and ’Temperature’
mae_rmse_results = []

for start in range(start_year, end_year, group_size):
    df_group = df[(df[’Year’] >= start) & (df[’Year’] < start +
        group_size)].copy()
    print(f"\nAnalyzing period: {start}-{start + group_size}")

    # Decompose time series
    decomposition = seasonal_decompose(df_group[’Temperature’], model=
        ’additive’, period=365)
    fig, (ax1, ax2, ax3) = plt.subplots(3, 1, figsize=(14, 8))
    decomposition.trend.plot(ax=ax1, title=’Trend’)
    decomposition.seasonal.plot(ax=ax2, title=’Seasonality’)
    decomposition.resid.plot(ax=ax3, title=’Residuals’)
    plt.tight_layout()
    plt.show()

    # Normalize Data
    scaler = MinMaxScaler(feature_range=(0, 1))
    df_group[’Temperature_Scaled’] = scaler.fit_transform(df_group[[’
        Temperature’]])

    # Prepare Sequences for LSTM
    seq_length = 30  # Using 30 days of past data
    X, y = create_sequences(df_group[’Temperature_Scaled’].values,
        seq_length)

    # Reshape for LSTM
    X = X.reshape((X.shape[0], X.shape[1], 1))

    # Train-Test Split
    X_train, X_test, y_train, y_test = train_test_split(X, y,
        test_size=0.2, shuffle=False)

    # Build LSTM Model
    model = Sequential([
        LSTM(50, activation=’relu’, return_sequences=True, input_shape
            =(seq_length, 1)),
        LSTM(50, activation=’relu’),
        Dense(1)
    ])
    model.compile(optimizer='adam', loss='mse')

    # Train Model
    model.fit(X_train, y_train, epochs=20, batch_size=16, verbose=1)

    # Predictions
    y_pred_lstm = model.predict(X_test)
    y_pred_lstm = scaler.inverse_transform(y_pred_lstm)

    # Evaluate Performance
    y_test_actual = scaler.inverse_transform(y_test.reshape(-1, 1))
    mae_lstm = mean_absolute_error(y_test_actual, y_pred_lstm)
    rmse_lstm = np.sqrt(mean_squared_error(y_test_actual, y_pred_lstm)
       )

    # Store Results
    mae_rmse_results.append([f"{start}-{start + group_size}", mae_lstm
       , rmse_lstm])
    print(f'LSTM MAE ({start}-{start + group_size}): {mae_lstm:.4f},
       RMSE: {rmse_lstm:.4f}')

# Create DataFrame for Results
results_df = pd.DataFrame(mae_rmse_results, columns=["Time Period", "
   MAE", "RMSE"])
print("\nSummary of LSTM Performance:")
print(results_df)

# Plot Results
plt.figure(figsize=(12, 6))
plt.plot(results_df["Time Period"], results_df["MAE"], marker='o',
   label="MAE")
plt.plot(results_df["Time Period"], results_df["RMSE"], marker='s',
   label="RMSE")
plt.xlabel("Time Period")
plt.ylabel("Error Value")
plt.title("MAE and RMSE for Each 10-Year Group using LSTM")
plt.xticks(rotation=45)
plt.legend()
plt.tight_layout()
plt.show()
```

Listing 5.3: Python Code: LSTM Modelling

Bibliography

Foster, G. and Rahmstorf, S. (2011a). Global temperature evolution 1979–2010. *Environmental research letters*, 6(4):044022.

Foster, G. and Rahmstorf, S. (2011b). Global temperature evolution 1979–2010. *Environmental Research Letters*, 6(4):044022.

Hansen, J., Ruedy, R., Sato, M., and Lo, K. (2010). Global surface temperature change. *Reviews of Geophysics*, 48:RG4004.

Hansen, J., Sato, M., Ruedy, R., Lo, K., Lea, D. W., and Medina-Elizade, M. (2006). Global temperature change. *Proceedings of the National Academy of Sciences*, 103(39):14288–14293.

Lindsey, R. and Dahlman, L. (2020). Climate change: Global temperature. *Climate. gov*, 16:1–5.

Rignot, E., Casassa, G., Gogineni, S., Krabill, W., Martins, R., Mouginot, J., and Scheuchl, B. (2019). Four decades of antarctic ice sheet mass balance from 1979-2017. *Nature Geoscience*, 12(3):174–178.

Rohde, R. A. and Hausfather, Z. (2020a). The berkeley earth land/ocean temperature record. *Earth System Science Data*, 12(4):3469–3479.

Rohde, R. A. and Hausfather, Z. (2020b). The berkeley earth land/ocean temperature record. *Earth System Science Data Discussions*, 2020:1–16.

The Indian Express (2025a). 2024 confirmed as first year to breach 1.5°c threshold. *The Indian Express.* Accessed May 8, 2025.

The Indian Express (2025b). Nature study links glacier melt to rising sea levels. *The Indian Express.* Accessed May 8, 2025.

The Indian Express (2025c). Sea ice at record lows in both polar regions: Nsidc, bbc analysis. *The Indian Express.* Accessed May 8, 2025.